Institute for Molecular Medicine Finland

University of Helsinki

Helsinki, Finland

Association and interplay of genetic and epigenetic variants in smoking behavior

Richa Gupta

ACADEMIC DISSERTATION

To be presented, with the permission of the Faculty of Medicine of the University of Helsinki, for public examination in Seminar room 3, Biomedicum 1 Helsinki

on 20 April 2018 at noon.

Helsinki 2018

Cover layout by Anita Tienhaara

Cover picture by Nikita Mathur

ISBN 978-951-51-4134-7 (paperback)

ISBN 978-951-51-4135-4 (PDF)

Unigrafia Oy

Helsinki 2018

Supervisors

Jaakko Kaprio, MD PhD
Professor in genetic epidemiology
Institute for Molecular Medicine Finland (FIMM)
University of Helsinki, Helsinki, Finland

Miina Ollikainen, PhD
Adjunct Professor in epigenetics
Institute for Molecular Medicine Finland (FIMM)
University of Helsinki, Helsinki, Finland

Reviewers

Harri Lähdesmäki, PhD
Associate Professor at the Department of Computer Science
Aalto University School of Science, Espoo, Finland

Eilis Hannon, PhD
Bioinformatics Research Fellow at University of Exeter Medical School
University of Exeter, Exeter, UK

Opponent

Bas Heijmans, PhD
Associate Professor at the Department of Molecular Epidemiology
Leiden University Medical Center, Leiden, Netherlands

“The more I learn, the more I realize how much I don't know”

– Albert Einstein

Abstract

Smoking is a major preventable risk factor for global morbidity and mortality and is influenced by both genetic and environmental factors. With the central goal of understanding the links between the complex trait of smoking behavior and genetic as well as epigenetic variation in the genome, this thesis focuses on identifying novel associations and validating the involvement of candidate genes in smoking behavior. An unbiased, hypothesis-free genome-wide scanning approach for association with genetic variants (single nucleotide polymorphisms; SNPs) in study I and epigenetic variants (DNA methylation) in study II were applied utilizing biomarkers of nicotine metabolism and exposure respectively. Taking a hypothesis-driven targeted approach, in study III, the involvement of neuregulin signaling pathway genes in smoking behavior was examined and validated in a phenotypically rich family sample. With the increasing and due emphasis on interpretation of associations identified in non-protein coding parts of the genome, we investigated the regulatory potential of the highlighted variants as well as assessing mediation via epigenetic mechanisms, utilizing in-house data and publicly available multi-omics resources.

In study I, utilizing a biomarker for nicotine clearance rate (nicotine metabolite ratio, NMR) in a genome-wide association study (GWAS) meta-analysis, we identified novel association on chromosome 19, represented by three independent loci mapping to or near the main nicotine metabolizing enzyme *CYP2A6*. We examined the regulatory potential of the SNPs and identified a subset of genome-wide significant SNPs (173 of 719) as methylation quantitative trait loci (meQTLs). Using causal inference test, we observed methylation at one CpG in *DLL3* mediates the effect of genotype on NMR. We also constructed a genetic risk score (GRS) with the independent SNPs identified in the GWAS meta-analysis, which explain ~30% variance observed in NMR. As evidenced by clinical trial studies, NMR influences the efficacy of cessation pharmacotherapeutics highlighting the potential value of our findings.

In study II, we utilized cotinine, an established biomarker of nicotine exposure, and performed epigenome-wide association study (EWAS) in regular smokers. We identified several novel loci in smoking-related genes, underlining the value of using biological phenotypes. Cotinine levels are influenced by nicotine intake as well as nicotine clearance, we utilized the GRS constructed in study I to account for such confounding, identifying additional novel loci. We further assessed the role of genetic variants in the highlighted genes and identified several *cis* and *trans* meQTLs. A handful of these meQTLs were also directly associated with cotinine levels.

At these loci, we examined whether DNA methylation is a mediator between the observed association of genotype and cotinine levels and detected mediation at seven CpG sites, implying DNA methylation may be a cause, not a consequence of nicotine exposure, as commonly assumed. Our results point at an interplay between the genome and epigenome while identifying novel nicotine exposure pertinent loci.

In study III, we applied a targeted approach to examine the role of the neuregulin signaling pathway (NSP) genes in smoking behavior utilizing a phenotypically rich family sample. Extensive association and joint linkage and linkage disequilibrium analysis of common, low frequency and rare variants in the ten key NSP genes with a wide spectrum of smoking behavior phenotypes revealed significant association with seven of the ten genes. No significant associations were observed with alcohol use phenotypes hinting at NSP's involvement specifically in smoking instead of addictions in general. Utilizing integrative methods and multi-omics data from an independent population sample and publicly available resources, we show the majority (56 of 66) of highlighted SNPs have regulatory potential. Our results provide evidence for the involvement of the NSP in smoking behavior, a candidate pathway for smoking and comorbid disorders.

Key points from this thesis:

1. Biomarkers can be powerful in identifying meaningful associations even with moderate sample sizes in complex behavioral traits.
2. Genetic and epigenetic differences between individuals influence smoking behavior via genes involved in nicotine metabolism, nicotine dependence, and neuronal pathways.
3. DNA methylation is a molecular mechanism mediating the effects of genotype on smoking behavior phenotypes at some loci.
4. Multi-omics data including, but not limited to, genetics, epigenetics, and transcriptomics can immensely aid in assessing the functional consequences of otherwise seemingly non-functional associations identified via genome-wide association studies providing potentially druggable targets for personalized medication.

Table of Contents

List of original publications

This thesis is based on the following publications:

I. Loukola A, Buchwald J, **Gupta R**, Palviainen T, Hällfors J, Tikkanen E, Korhonen T, Ollikainen M, Sarin AP, Ripatti S, Lehtimäki T, Raitakari O, Salomaa V, Rose RJ, Tyndale RF, Kaprio J. *A Genome-Wide Association Study of a Biomarker of Nicotine Metabolism.* PLoS Genet. 2015 Sep 25;11(9):e1005498. doi: 10.1371/journal.pgen.1005498. PubMed PMID: 26407342

II. **Gupta R**, Tyndale R, Velagapudi V, Korhonen T, Kaprio J, Loukola A, Ollikainen M. *Epigenome-wide association study of serum cotinine in regular smokers reveals novel genetically driven loci*. Submitted.

III. **Gupta R**, Qaiser B, He L, Hiekkalinna TS, Zheutlin AB, Therman S, Ollikainen M, Ripatti S, Perola M, Salomaa V, Milani L, Cannon TD, Madden PAF, Korhonen T, Kaprio J, Loukola A. *Neuregulin signaling pathway in smoking behavior*. Transl Psychiatry. 2017 Aug 22;7(8):e1212. doi: 10.1038/tp.2017.183. PubMed PMID: 28892072.

The publications are referred to in the text by their roman numerals. All publications are reprinted at the end of this book with permissions (where applicable) from the publishers.

Abbreviations

CIT	Causal Inference Test
DILGOM	DIetary, Lifestyle and Genetic determinants of Obesity and Metabolic syndrome
DSM-IV	Diagnostic and Statistical Manual of Mental Disorders, 4th Edition
EWAS	Epigenome-Wide Association Study
eQTL	Expression Quantitative Trait Loci
FTC	Finnish Twin cohort
GRS	Genetic Risk Score
GWAS	Genome-Wide Association Study
LD	Linkage Disequilibrium
meQTL	Methylation Quantitative Trait Loci
MAF	Minor Allele Frequency
NAG-FIN	Nicotine Addiction Genetics - Finland
ND	Nicotine Dependence
NMR	Nicotine Metabolite Ratio
NSP	Neuregulin Signaling Pathway
NW	Nicotine Withdrawal
PC	Principal Component
QTL	Quantitative Trait Loci
SCZ	Schizophrenia
SI	Smoking Initiation
SNP	Single Nucleotide Polymorphism
QC	Quality Control

1 Introduction

All of us are 99.9% genetically identical, yet phenotypically quite different. These differences result from not only the small proportion of differences in our genetic code but also from the environmental influences on the function of our genome. Influenced by both genes and the environment, these interindividual differences hold secrets to many complex diseases that burden humanity with morbidity and mortality. As such, a better understanding of the underlying mechanisms involved in complex diseases can aid in finding effective treatments.

Identifying susceptibility genomic loci underlying complex traits using genome-wide association analysis has been successful and continues to be the most popular means for screening the genome. However, the majority of the identified associations reside in non-coding regions of the genome. Unlike the variants located in protein-coding parts of the gene, which may alter the amino acid sequence and consequent protein product, variants in non-coding regions have an unclear regulatory function. Lately, the focus has shifted towards integration of multiple biological data layers (Multiple-omics) such as genetic, epigenomic and transcriptomic, to facilitate a more comprehensive understanding of the otherwise seemingly non-functional associations.

This thesis covers association and interplay of genetic and epigenetic variants in the context of smoking behavior, a major preventable risk factor of global morbidity and mortality. Genetics assessed by single nucleotide polymorphisms (SNPs), epigenetics assessed by DNA methylation, as well as transcriptomic data were utilized to further infer functional implications and consequences of associations identified. Figure 1 is a pictorial representation of the work included in the thesis showing the approaches employed, piecing together the complex biological picture of smoking behavior.

Genome-wide association study (GWAS), a well-established means of identifying genetic variants associated with complex traits, was applied in study I to identify genomic regions associated with the rate of nicotine metabolism, a major influencer of smoking behavior. For instance, to maintain nicotine levels, fast metabolizers smoke more due to quicker clearance of nicotine, whereas slow metabolizers smoke less often as nicotine levels are maintained for longer from a given intake. Taking a similar hypothesis-free approach to scan the epigenome, an epigenome-wide association study (EWAS) was applied to assess the association of serum cotinine, a reliable biomarker of nicotine exposure, in study II to identify genomic loci that might

influence smoking behavior by epigenetic mechanisms. In study III, hypothesis-driven targeted (candidate gene) approach was used to assess the role of neuregulin signaling pathway genes, with a wide spectrum of phenotypes encompassing smoking behavior and alcohol use and abuse.

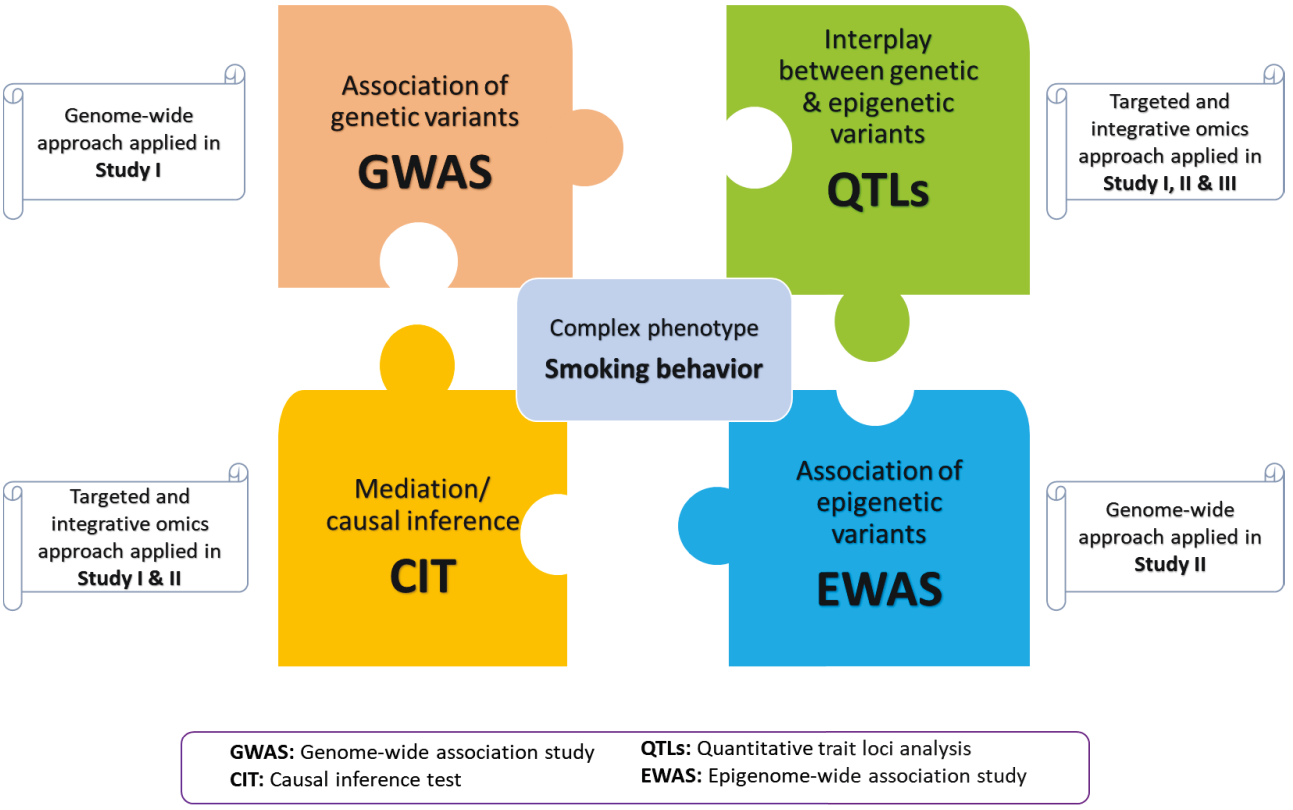


Figure 1. *Methodologies applied in the studies included in this thesis to uncover underlying biology of the complex behavioral trait of smoking.*

To uncover the regulatory potential of variants associated with smoking behavior phenotypes, integrative analysis using expression and DNA methylation data were employed in all three studies. The role of DNA methylation as a molecular mechanism mediating the effects of the observed association between genotype and phenotype was also examined in study I and II. In the next chapter, literature review covering what is known and what we don't about the genetics and epigenetics of smoking behavior is presented, followed by aims of the three studies. Next chapter provides a description of materials and methods employed, followed by the main findings and discussion from the three studies. Final chapters present challenges, future prospects, and conclusions.

2 Literature Review

Several factors are involved in the development and progression of human diseases and traits. Complex traits influenced by both genes and the environment may be expressed differently between individuals. The advancement in technology and reduction in associated costs has led to an extensive characterization of the human genome consequently increasing our understanding of the underlying biological mechanisms of many complex traits. The contribution of genetic factors (quantifiable by the advent of technology) *versus* environmental factors (difficult to quantify) vary across phenotypes and determining a complete picture remains challenging. Table 1 provides a glossary of terms that will be used throughout this thesis.

Table 1. *Glossary*

Biomarker	A biological identifier characteristic of a disease, pathological or physiological process etc. In this thesis, for example, metabolites of nicotine are used as biomarkers of nicotine exposure and nicotine metabolism.
CpG	Site in the genome where nucleotide cytosine is adjacent to guanine; DNA methylation usually occurs at these sites.
Epigenome	Chemical compounds that overlay the genome and regulate gene activity without altering genomic DNA sequence. DNA methylation is the most commonly studied epigenetic mark.
EWAS	Epigenome-Wide Association Study; the study of the association between epigenetic variation (in this thesis, DNA methylation) to a trait or disease.
Genome	The complete set of genetic information sequence.
GWAS	Genome-Wide Association Study; the study of the association between genetic variation (SNPs) in the genome to a trait or disease.
Omics	A wide spectrum of technologies used to explore the roles, relationships, and functions of the various biological molecules in an organism.
Phenotype	The trait of interest; set of observed characteristics of an individual arising from the interplay between the genome and the environment. For example, behavior, disease etc.
QTL	Quantitative Trait Loci; a locus in the genome associated with a quantitative trait (molecular measures). For example, gene expression (eQTL) and/or DNA methylation (meQTL)
SNP	Single nucleotide polymorphism; variable site in the genome.

2.1 Genetic variation and heritability of traits

The human genome has more than 3 billion base pairs. Genetic variation such as single nucleotide polymorphism (SNPs), insertions, deletions, and copy number variation exists in the genome and contribute toward interindividual differences. The most common form of genetic variation are the SNPs, occurring at every 300 nucleotides on average, meaning about 10 million SNPs exist in the genome. SNPs occur predominantly in the non-coding parts of the genome and have variable occurrence (allele frequencies) across populations.

Twin and family studies have indicated the contribution of genetics to almost every aspect of life [1, 2]. This variance observed in a trait attributable to genetics is called heritability [3] and has been used as a basis for identifying genomic loci associated with complex traits. One popular means of examining the association of a phenotype with genetic variation (SNPs) is genome-wide association studies (GWAS). GWAS have been crucial and successful in identifying genetic loci targeting a multitude of biological pathways associated with complex traits [4]. The GWAS catalog, reporting over 50,000 unique SNP-trait associations (as of September 2017) is a testament to the success GWAS approach has seen in the past decade [5]. As such, hypothesis-free GWAS approach, along with hypothesis-driven candidate gene approach has identified genetic susceptibility underlying several complex traits [6, 7]. For instance, in a genetic association study of smoking behaviors [8], the chr15q25 region spanning the nicotinic receptor genes was identified in a GWAS of smoking quantity. In the same study [8], hypothesis based candidate gene approach provided evidence of association in the monoamine oxidases gene, a contributor towards reinforcing and motivating effects of smoking [9].

Trait-associated SNPs identified by GWAS explain less than 50% of estimated heritability [10, 11], with a substantial proportion remaining unexplained. Yet, the effect of multiple SNPs identified via GWAS summed up into a genetic risk score (GRS), has been used as a measure of genetic susceptibility and holds great utility, as a proxy for the trait in different samples [12, 13]. Common and low-frequency variants explain a fairly small proportion of the genetic contribution to variance observed in common traits [14]. It has been long speculated that rare variants (with large effects) and other structural variations (insertions, deletions, and copy number variation), variants not tagged by the arrays, non-additive genetic effects, and gene-environment interactions may explain some of the missing heritability [15]. Rare variant effects can be captured using family samples and population isolates [16], while sequencing technology can help identify variants undetectable by microarrays, adding additional information toward missing

heritability. Lately, epigenetics has also come to focus and is considered a valuable contributor toward missing heritability [17].

2.2 Epigenetics and its contributions in trait variation

Epigenetics, as defined in Table 1, refers to a set of chemical modifications that play a key role in regulating gene expression by altering DNA accessibility and chromatin structure [18, 19]. Epigenetic mechanisms include DNA (methylation) and histone modifications (methylation and acetylation), as well as regulation by non-coding RNAs [20]. High-risk genetic variants identified in association studies reside in regulatory regions of the genes [21] which are the main sites of epigenetic regulation. Therefore, mapping epigenetic changes may shed some light on the missing heritability. The human epigenome has been intensively studied in the last decade, with large-scale projects like the Roadmap, ENCODE and blueprint projects [22-24] mapping regulatory features across the genome. Most commonly studied epigenetic alteration is DNA methylation (referred to as methylation here on in the thesis), wherein a methyl group is attached to cytosine of cytosine-guanine dinucleotides (CpG site; Table 1) or non-CpG sites such as CHG or CHH (where H correspond to A, T or C). However, methylation is almost exclusively found in CpG dinucleotides. This dynamic epigenomic feature is tissue and cell type specific and is influenced by environmental factors like age and lifestyle (for example - diet, smoking, and drinking habits). Aside from these factors, methylation variation is also largely driven by genetic variation [25]. Although conventionally viewed as a transcriptional repressor, methylation has been associated with increased transcriptional activity as well [19] and is one of the mechanisms by which gene expression is regulated. Like GWAS, unbiased screening of the epigenome with epigenome-wide association studies (EWAS) is gaining importance in identifying novel disease-associated genomic loci that may be under epigenetic regulation and are not discovered through genetic studies [26].

2.3 Regulatory potential of trait-associated variants

As mentioned above, association studies have been successful in identifying genomic regions associated with complex traits and diseases. However, it is difficult to figure out the causative variant underlying each association signal and its effect in most cases. Most of the genetic associations identified in GWAS, not surprisingly, reside in the non-coding region of the genome. Unlike variants in the protein-coding parts of the genome, where the genetic variation may function by altering the final protein product of the gene, variants in the non-coding parts

of the genome can pose difficulty in inferring their functional role. Regulatory potential of such variants has taken the focus of current research [27-31].

There is growing evidence for the contribution of genetic variants toward gene expression and methylation patterns [32]. Trait-associated variations in the DNA sequence could affect gene function via alteration of expression or methylation, or any other epigenetic mechanism. A genetic locus that affects a molecular trait such as gene expression, methylation or histone modifications, is called a quantitative trait loci (QTL). Figure 2 illustrates how molecular data layers, such as epigenetic and transcriptomic, can be useful in identifying the mechanistic role of the variants identified by genetic association studies, providing greater insight into the consequential effects of such variants [33]. These hereditary traits also provide a plausible mechanism by which methylation and expression patterns could be different under environmental exposures.

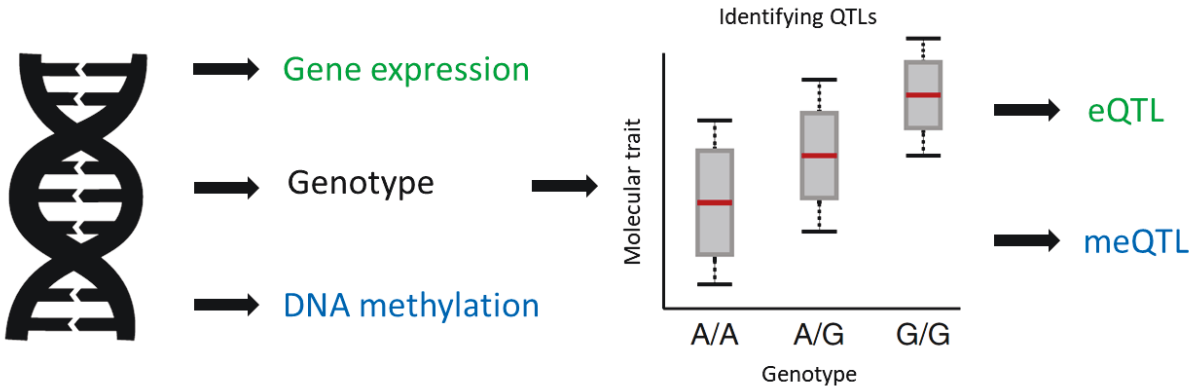


Figure 2. *Molecular data including the genetic sequence, gene expression and methylation from the tissue of interest once obtained, can be utilized to assess whether the genetic variants are quantitative trait loci (QTL). An observed association between the genotype and molecular traits such as expression (eQTL) and DNA methylation (meQTL) may explain how the trait associated genetic variants may regulate the phenotype of interest.*

The QTLs provide valuable information on whether or how the genetic variation may be regulating gene function. It would be of further interest to explore mediation via therapeutically targetable mediators, for example, methylation. With multiple layers of biological data (multi-omics; see section 2.4 below), mediation via molecular mechanisms between trait and genotype can be tested using statistical frameworks such as causal inference test and Mendelian randomization [34, 35]. This is tremendously useful in identifying mediators which provide additional targets for therapeutic intervention.

Integration of multiple layers of biological data can bring more information than analysis of single layers alone. The ability to combine multiple layers of molecular data with each other will allow uncovering further systemic information regarding biological modifications that occur in complex traits. The application of this approach has the potential of identifying targets for therapeutic interventions.

2.4 The era of omics - quantifiable biological data

Omics, referring to different layers of biological data aims at exploring the functions of as many genes as possible. The advancement in technology, especially the low-cost microarray, has powered tremendous progress in genomic research. Figure 3, provides an overview of the workflow of a microarray. The basic workflow remains unvarying across different omics data generated by microarray with differences present in sample preparation and array designs (oligobeads *versus* features spotted on to the array).

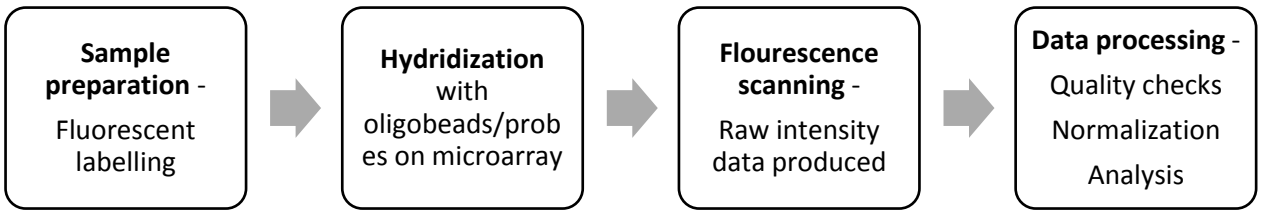


Figure 3. *This figure represents a general workflow of microarray chip technology with differences present at each step between genotype, DNA methylation, and gene expression array designs and chemistry.*

Large-scale sequencing projects such as the 1000 genomes [36], HapMap [37], and Haplotype Reference Consortium [38] have enabled cost-effective genotyping of individuals on large-scale. For example, genotyping can be performed with SNP microarray that tags only a handful of SNPs and genomic reference panels based on the large-scale sequencing projects can then be used for imputation of unobserved (not directly genotyped) variants providing a broad view of the genome. In capturing methylation using microarray a key step is bisulfite conversion. Treating denatured DNA with sodium bisulfite helps in distinguishing methylated *versus* unmethylated cytosine. Bisulfite conversion leads to deamination of unmethylated cytosine residues to uracil, leaving methylated cytosine intact. The uracil is subsequently amplified to thymine and detection of cytosine and thymine, like in genotyping arrays, depicts the methylation status at a CpG site. For measuring the expression of genes, messenger RNA isolated from the samples is labeled with fluorescent dye and reverse transcribed to generate

complementary DNA sequence which is hybridized to the microarray. The fluorescence measure reveals the relative expression of genes across samples.

With the early development and popularity of microarray technology, standard protocols for data processing are well established for genotyping and gene expression analysis. However, methylation arrays being relatively new and owing to a different design - two types of probes on the same array (Infinium Human Methylation 450K and EPIC arrays), still lack consensus on a standard protocol for data processing. Numerous pre-processing and normalization methods for methylation array have been implemented and comprehensively reviewed [39-48]. With each method having limitations and advantages, a consensus has been difficult to reach [49]. Quantile normalization is performed to transform all the arrays to have a common distribution of intensities [50]. However, the two-probe biochemistry, although greatly increases genomic coverage, necessitated further within-array normalization methods to be developed such as Beta-mixture quantile dilation [51], Subset-quantile within array [52] and dasen [47], which allow the signals from both probe types to be analyzed together. Other normalization methods such as Stratified quantile normalization [53] took into consideration the genomic localization of probes to adjust for technical variation. In a recent study by Lehne et al. [54], quantile normalization of raw intensity values categorized by probe type, color channel, and probe subtypes (similar in concept to functional normalization [55]) outperformed existing normalization methods and is gaining acceptance even for large-scale consortia analysis.

2.5 Smoking behavior

Among environmental factors that adversely affect the health is tobacco smoking, a major avoidable source of mortality across the globe [56]. Smoking is a risk factor for dire diseases like cancer, cardiovascular and respiratory disorders [57]. Smoking behavior is a complex phenotype entailing different aspects including, but not limited to, smoking initiation, nicotine dependence, nicotine withdrawal, and cessation [12]. There are several factors that influence smoking initiation, for example, peer pressure in adolescence [58], gender, familial conditions (socioeconomic status and habits of other members), ethnicity, and other substance use [59-61]. Not all individuals who initiate smoking develop nicotine dependence. However, smokers who develop nicotine dependence due to persistent smoking might find it difficult to quit despite knowing the adverse consequences. Nicotine, the addictive component of cigarette, has a positive reinforcing effect and is one of the key factors determining smoking cessation success [57]. Nicotine withdrawal symptoms such as anxiety, irritability, concentration problems,

depressed mood, and cravings make abstinence difficult consequently resulting in high relapse rates [62, 63]. Different aspects of smoking behavior can be captured with self-report questionnaires, objective assessment measures such as the Diagnostic and Statistical Manual of Mental Disorders (DSM) [64] and the Fagerström Test of Nicotine Dependence (FTND) [65] and biochemical measures such as cotinine levels in urine, saliva or serum can quantify nicotine exposure [66]. It is worth noting that each form of measure taps somewhat overlapping yet distinct domains of smoking behavior. For instance, DSM-IV based diagnosis and FTND are both measures of nicotine dependence, however, FTND is more reflective of smoking heaviness (higher correlation with both self-report and biochemical measures) and cessation likelihood than DSM based nicotine dependence [67].

Nicotine metabolism. One of the key drivers of nicotine dependence is the addictive potential of nicotine, hence understanding the mechanisms involved in nicotine breakdown in the body are of great importance. Nicotine is metabolized mainly by cytochrome P450 oxidases (primarily *CYP2A6*) in the liver [68], with ~75% of nicotine being converted to cotinine, which is further metabolized to *trans* 3-hyrdoxycotinine [69]. The ratio of 3-hydroxycotinine to cotinine, called the nicotine metabolite ratio (NMR), is an established biomarker of *CYP2A6* activity and can be used as a proxy for nicotine clearance rate. NMR critically affects smoking behavior, for example, individuals with faster nicotine metabolism tend to smoke more in order to self-titrate [70], putting them at higher risk of developing smoking-associated diseases [71]. On the other hand, individuals with slower metabolism tend to smoke less often as nicotine from a given intake is maintained for longer in their body. Such difference affects how much and how often a person smokes.

Cotinine, the primary metabolite of nicotine, has a long half-life (15–20 hours) [66] and is a reliable indicator of recent nicotine exposure, making it superior to the self-reported smoking quantity which is error-prone due to misreporting and recall bias. Cotinine levels are affected by nicotine intake i.e. how much a person smokes. In addition, it is also influenced by the rate of nicotine metabolism i.e. formation of cotinine from nicotine as well as clearance of cotinine (converted to *trans* 3-hydroxycotinine). This metabolism is largely mediated by *CYP2A6* and can be measured with NMR. As described above, NMR also affects smoking behavior by altering nicotine intake, making its influence on cotinine levels more prominent. Metabolites have been highly successful in identifying genetic susceptibility loci [72, 73] as they provide biological proximity as well measurement precision. Biomarkers such as cotinine and NMR provide great

statistical power in association studies [74, 75]. In a GWAS meta-analysis of cotinine [74], aside from the most consistently identified genome-wide significant locus associated with smoking i.e. nicotinic receptor gene cluster, a locus on chromosome 4 encompassing *UGT2B10* gene was identified. This gene although involved in nicotine and cotinine metabolism [76] was not associated with smoking quantity [74], indicating that cotinine captures more than mere nicotine exposure. This is in line with the factors that influence cotinine levels described above.

Comorbidities. Smoking is highly comorbid with other behavioral phenotypes, the most prominent one being alcohol use [77]. Smoking is also associated with several neuropsychiatric disorders [78-80]. One such disorder is schizophrenia (SCZ), where one hypothesis is that the diseased tend to smoke more in order to self-medicate [81]. Such comorbidities further complicate smoking cessation. Many cessation pharmacotherapies such as nicotine replacement therapy, bupropion, and varenicline are available but show a maximum of two to three-fold success rate [82]. There is a high demand for smoking cessation therapeutics to aid smokers trying to quit as well as target comorbid illnesses.

2.5.1 Genetics of smoking

Smoking behavior is a multifactorial trait with substantial genetic influence [83], and its genetic contribution has been well reviewed [84, 85]. Heritability estimates vary across different aspects of smoking behavior and ranges between 35% - 55% for smoking initiation [86, 87], 31% - 75% for nicotine dependence [86, 88-90], 26%-50% for nicotine withdrawal [91-94] and have been reported as high as 81% for rate of nicotine metabolism [75, 95, 96]. In 2010, three large-scale GWAS meta-analysis (N~140,000) identified genome-wide significant loci underpinning different stages of smoking [97]. Associations highlighted the role of brain-derived neurotrophic factor gene (*BDNF*) in smoking initiation [98], nicotinic receptor gene cluster (*CHRNA5–CHRNA3–CHRNB4*) [98, 99] and *CYP2A6* in smoking dependence [100], and dopamine β-hydroxylase (*DBH)* gene with smoking cessation [98, 101]. Numerous other genetic studies have identified many other genetic loci associated with different aspects of smoking.

Nicotine Metabolite Ratio. As described above, NMR is a key influencer of smoking behavior and varies significantly between individuals [102, 103] with high heritability estimates (~80%), suggesting major genetic contribution to the variance observed in NMR [75]. Clinical trial studies stratifying participants on their NMR profiles indicated its influential role in the efficacy of smoking cessation pharmacotherapies [104, 105]. It is worth noting that only a fraction of interindividual differences in nicotine metabolism is accounted for by known reduced activity

CYP2A6 variants (www.pharmvar.org/htdocs/archive/cyp2a6.htm) [106] and up to 85% remains unexplained [107]. GWAS in different population samples have identified population-specific variants associated with NMR [75, 108, 109].

In addition to the genes identified by hypothesis-free screening of the genome, the role of pathways and candidate genes in distinct and overlapping stages of smoking progression has become evident [110, 111]. One such set of candidate genes comprise the main functional components of the neuregulin signaling pathway (NSP).

Neuregulin signaling pathway. The NSP, a modulator of neuronal migration and differentiation has ten key functional components presented as a protein-protein interaction (PPI) network in Figure 4. Table 2 further provides the functions of the ten genes. The NSP has been implicated in nicotine dependence and associated psychiatric comorbidities [112].

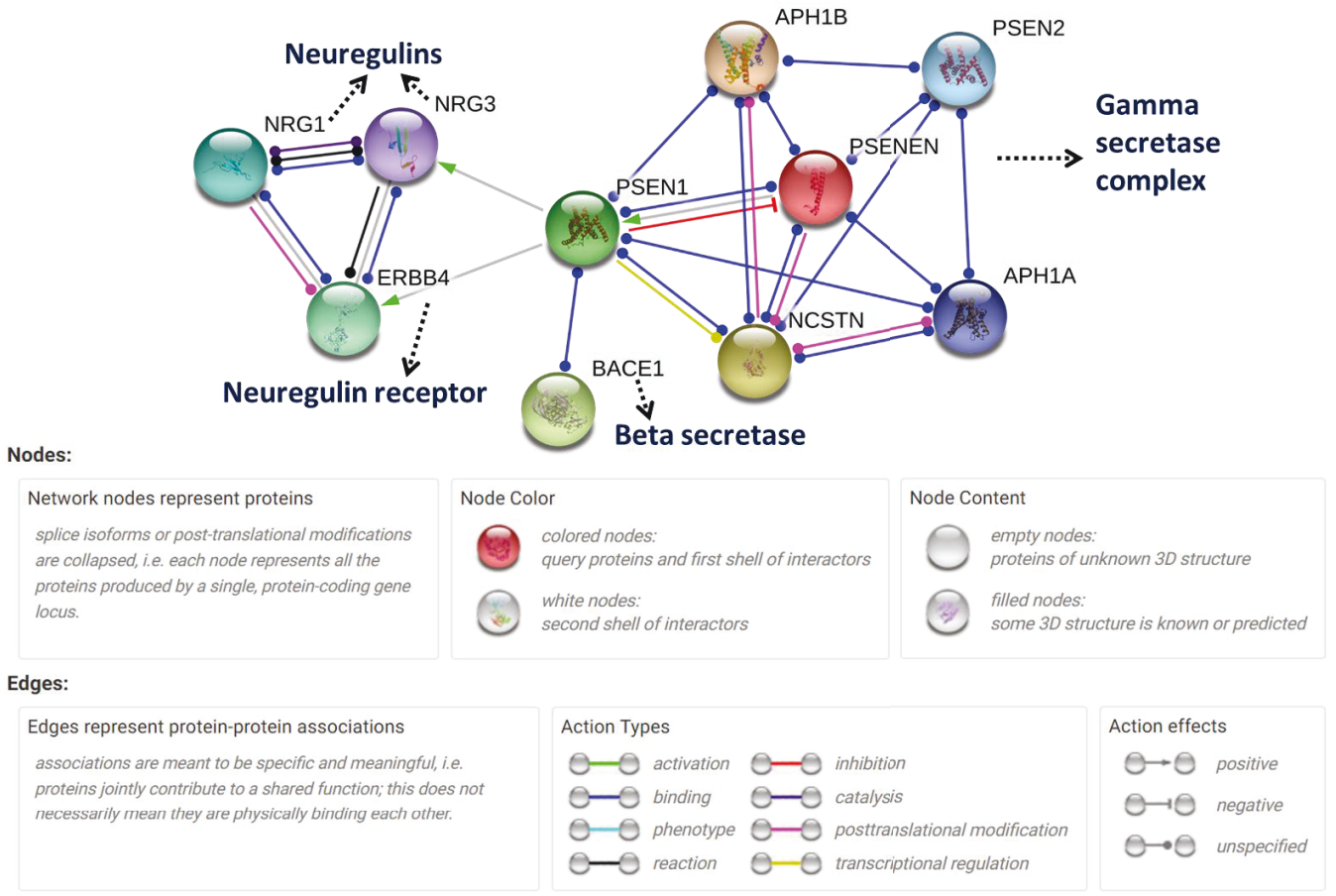


Figure 4. *The neuregulin signaling pathway as a protein-protein interaction (PPI) network. The key components of the NSP include the neuregulin signaling molecules (NRG1 and NRG3), their receptor (ERBB4), beta-secretase (BACE1) and gamma-secretase complex (PSEN1, PSEN2, APH1A, APH1B, PSENEN, and NCSTN). Modified figure produced using string-db.org [113], a PPI network database.*

Table 2. *Summary of the ten key genes comprising the neuregulin signaling pathway. Gene summaries derived from STRING database [113] in conjunction with Figure 4.*

Gene	Chr	Function
Neuregulins		
NRG1	8	Neuregulin 1; Direct ligand *ERBB4* tyrosine kinase receptors. Concomitantly recruits *ERBB1* and *ERBB2* coreceptors, resulting in ligand-stimulated tyrosine phosphorylation and activation of the ERBB receptors. The multiple isoforms perform diverse functions such as inducing growth and differentiation of epithelial, glial, neuronal, and skeletal muscle cells; inducing expression of acetylcholine receptor in synaptic vesicles during the formation of the neuromuscular junction.
NRG3	10	Neuregulin 3; Direct ligand for the *ERBB4* tyrosine kinase receptor. Binding results in ligand-stimulated tyrosine phosphorylation and activation of the receptor.
Neuregulin receptor		
ERBB4	2	V-erb-a erythroblastic leukemia viral oncogene homolog 4 (avian); Tyrosine-protein kinase that plays an essential role as cell surface receptor for neuregulins and EGF family members and regulates development of the heart, the central nervous system and the mammary gland, gene transcription, cell proliferation, differentiation, migration and apoptosis
Beta secretase		
BACE1	11	Beta-site APP-cleaving enzyme 1; Responsible for the proteolytic processing of the amyloid precursor protein (APP). Cleaves at the N-terminus of the A-beta peptide sequence, leads to the generation and extracellular release of beta-cleaved soluble APP, and a corresponding cell-associated C-terminal fragment which is later released by gamma-secretase
Gamma secretase complex		
APH1A	1	Anterior pharynx defective 1 homolog A; Essential subunit of the gamma-secretase complex, an endoprotease complex that catalyzes the intramembrane cleavage of integral proteins. It probably represents a stabilizing cofactor for the presenilin homodimer that promotes the formation of a stable complex
NCSTN	1	Nicastrin; Essential subunit of the gamma-secretase complex, an endoprotease complex that catalyzes the intramembrane cleavage of integral membrane proteins. It probably represents a stabilizing cofactor required for the assembly of the gamma-secretase complex
PSEN2	1	Presenilin 2; Probable catalytic subunit of the gamma-secretase complex, an endoprotease complex that catalyzes the intramembrane cleavage of integral membrane proteins. Requires the other members of the gamma-secretase complex to have a protease activity. May play a role in intracellular signaling and gene expression or in linking chromatin to the nuclear membrane and may function in the cytoplasmic partitioning of proteins
PSEN1	14	Presenilin 1; Probable catalytic subunit of the gamma-secretase complex, an endoprotease complex that catalyzes the intramembrane cleavage of integral membrane proteins. Requires the other members of the gamma-secretase complex to have a protease activity. May play a role in intracellular signaling and gene expression or in linking chromatin to the nuclear membrane.
APH1B	15	Anterior pharynx defective 1 homolog B; Probable subunit of the gamma-secretase complex, an endoprotease complex that catalyzes the intramembrane cleavage of integral proteins. It probably represents a stabilizing cofactor for the presenilin homodimer that promotes the formation of a stable complex. Probably present in a minority of gamma-secretase complexes compared to *APH1A*
PSENEN	19	Presenilin enhancer 2 homolog; Essential subunit of the gamma-secretase complex, an endoprotease complex that catalyzes the intramembrane cleavage of integral membrane proteins. Probably represents the last step of maturation of gamma-secretase, facilitating endoproteolysis of presenilin and conferring gamma-secretase activity

One of the potential mechanisms by which NSP may be involved in nicotine dependence is through modulation of nicotinic acetylcholine receptor expression, which can be induced by binding of nicotine causing a cascading effect of conformational changes resulting in increased expression of *NRG3* [112]. The *NRG3* can be proteolytically cleaved by *BACE1* (beta-secretase) and bind to the *ERBB4* receptor or alternatively, *NRG3* may be cleaved by a gamma-secretase complex and can subsequently regulate gene transcription. The signaling molecule *NRG3* and its receptor *ERBB4* are associated with smoking initiation, nicotine dependence, and nicotine withdrawal symptoms in human and behavioral mouse model studies [114-116]. *NRG1* has been robustly associated with schizophrenia, a neurodevelopmental disorder with extremely high co-occurrence with smoking and nicotine dependence [81, 117]. Further, deficiency of *BACE1* or *APH1B* is reported to cause neurobehavioral abnormalities like schizophrenia in mice [118, 119]. Underlying genetic factors common to psychiatric disorders and smoking, make NSP a prime candidate for exploring the therapeutic potential to treat comorbidity observed between smoking, other substance use as well as psychiatric illness.

2.5.2 Epigenetics of smoking

Smoking has a major influence on methylation changes across the genome, as evidenced by numerous EWAS conducted in last few years and has been extensively reviewed [120, 121]. To highlight the incredible amount of evidence supporting the prominent effect of smoking on methylation, Table 3 summarizes the findings from EWAS conducted in healthy adults using methylation profiled in peripheral blood thus far (September 2017).

Table 3. *Table summarizing the findings from EWAS studies conducted in methylation profile of peripheral blood from healthy adults in chronological order.*

Reference	Assay*	Sample Size	Summary
Breitling et al. [122]	27k	177	First study examining genome-wide methylation in context of smoking in Caucasians. Only one CpG cg03636183 in *F2RL3* identified.
Wan et al. [123]	27k	1085	Two loci in *F2RL3* and *GPR15* identified. Authors suggested methylation responds to change in smoking behavior.
Siedlinski et al. [124]	Golden Gate	316	Only one CpG site in the *TGFBI* gene was associated with ever-smoking after adjusting for age and sex.
Shenker et al. [125]	450k	374	First report of association at *AHRR* with smoking exposure among other novel loci.
Philibert et al. [126]	450k	399	Association at *AHRR* locus even with low levels of smoking (less than half a pack-year) in a sample of young African Americans (average age 19 years)
Sun et al. [127]	27K	972	Loci in *F2RL3* and *GPR15* replicated in African Americans, consistent with findings in Caucasians

Zeilinger et al. [128]	450k	1793	Widespread effects of smoking observed in all chromosomes, with methylation at a few CpGs explaining upto 41% variance. Authors also suggested reversal of methylation levels upon smoking cessation.
Philibert et al. [129]	450k	107	First study using serum cotinine in a sample of young African American men. Only 2 CpG sites identified in *AHRR.*
Besingi et al. [130]	450k	432	Reported that methylation changes are caused by burnt products of tobacco based on their observation of no significant association between methylation and snuff (smokeless tobacco).
Elliot et al. [131]	450k	192	Reported ethnic differences (Europeans vs South Asians) in smoking associated methylation patterns. Authors also constructed a smoking score for prediction of smoking status.
Dogan et al. [132]	450k	111	Identified close to 1000 CpG sites associated with smoking in African American women.
Tsaprouni et al. [133]	450k	464	First report of a meQTL (rs2697768), affecting methylation levels at smoking associated cg03329539. Authors also reported this meQTL regulates expression of the *CHRND* (Cholinergic Receptor Nicotinic Delta Subunit) gene and that methylation levels are reversible upon cessation.
Flanagan et al. [134]	450k	92	Reported methylation pattern at the 2q37.1 and *AHRR* loci are stable over time as well as associated with time since quitting.
Guida et al. [135]	450k	745	Conducted in an all women population sample (European), the authors reported two types of CpG methylation patterns: ones that revert to levels of never smokers in less than a decade while others that do not, even after 35 years of smoking cessation.
Zaghlool et al. [136]	450k	123	The first study in Arab population, replicating findings from previous studies in Caucasians.
Allione et al. [137]	450k	40	Using a discordant twin pair design, the authors reported associations between methylation and smoking that are free of genetic confounding.
Qiu et al. [138]	27k	85	Examined association of CpG methylation with smoking and genetic variation in the vicinity of the highlighted CpG sites identifying meQTLs.
Sergi-Baixeras et al. [139]	450k	645	Among several other replicated loci, novel association was reported at a CpG site (cg06394460) in a smoking-associated gene *LNX2*.
Zhang et al. [140]	450k	500	Cotinine was used as a biomarker in EWAS. Authors also examined the use of methylation vs cotinine in distinguishing current from former and never smokers and reported methylation as a better measure.
Ambatipudi et al. [141]	450k	910	Identified several smoking associated loci and reported reversion of methylation levels at a subset of CpG sites upon cessation, while another subset of CpG sites did not respond to cessation even after 22 years like previously suggested by Guida et al. [135]
Joehanes et al. [142]	450k	15,907	First large meta-analysis including 16 cohorts identified methylation at ~7000 CpG sites related to smoking.
Lee et al. [143]	450k	100	First EWAS in the Korean population. Current smokers were verified using cotinine measures, however, cotinine levels were not used as a phenotype. Authors also report the association of gene expression in lung tissue with the top CpG sites.
Dogan et al. [144]	450k	1599	Authors report both distal and proximal interactions between genetic variants and smoking associated methylation; suggesting integrative analysis of epigenetic and genetic data is pivotal for a comprehensive understanding of the relationship between smoking and the genome.

***** *Methylation assessment platform used in each study.*

The most prominent associations observed consistently are in the aryl hydrocarbon receptor repressor (*AHRR*) and F2R Like Thrombin/Trypsin Receptor 3 (*F2RL3)* genes. *AHRR,* a key regulator of the relationships between the cell and the external environment, is involved in the xenobiotic metabolism such as polycyclic aromatic hydrocarbons (toxic components of cigarette smoke) [145]. *F2RL3* encodes the protease-activated receptor-4 which is likely involved in the pathophysiology of both cardiovascular and neoplastic diseases [146].

Methylation is tissue and cell-type specific. Although other smoking pertinent tissues such as buccal cells [147] and lungs [148] have also been studied, blood remains the primary tissue of choice because of the availability and accessibility. It should be noted that the methylation signature of smoking between buccal cells and blood although generally match for top ranking CpG sites, there was a ~40-fold higher association signal observed in buccal cells [147]. It is noteworthy that almost all the EWAS on smoking have utilized error-prone (misreporting or recall bias) self-reported measure of smoking, and the scope for utilizing a reliable source, such as a biomarker – Cotinine, would be valuable in capturing the direct effect of nicotine exposure on methylation. Smoking-related cardiovascular diseases [149, 150], cancer [151, 152], as well as maternal smoking during pregnancy and aberrant methylation in offspring [153] have also been extensively studied, implicating a wide-spread effect of smoking on the epigenome throughout the life course.

3 Aims

The focus of the work included in the thesis was to examine the association and interplay between genetic (SNPs) and epigenetic (DNA methylation) variants using hypothesis-free and hypothesis-driven approaches in smoking behavior. Specific aims for each of the studies included are listed below:

I. Identify novel genetic variants associated with the rate of nicotine metabolism using a biomarker and assess their functional consequences via epigenetic mechanisms.
II. Identify DNA methylation signature of nicotine exposure using serum cotinine levels in regular smokers. Investigate genetic contribution to the observed associations as well as assess whether methylation changes are a cause or consequence of nicotine exposure.
III. Validate the role of neuregulin signaling pathway genes in smoking behavior as opposed to addiction in general while examining the functional potential of the associating genetic variants using multi-omics data.

4 Materials and Methods

4.1 Study samples

In all three studies, data from Finnish population was analyzed. For replication of findings and functional annotation, publicly available data resources were utilized as detailed later in this section. Table 4 provides information on the samples and corresponding omics data analyzed within each cohort and study.

The Finnish Twin Cohort (Study I, II & III)

The Finnish twin cohort (FTC), was established in 1974 with the aim of studying genetic and environmental factors impacting complex diseases and associated behavioral risks. The FTC [154] has three main data sets including the Older Finnish Twins, FinnTwin16, and FinnTwin12, overviewed below:

1) **Older Finnish Twins (Study I, II & III).** Same-sex twin pairs (both monozygotic and dizygotic [155]) born before 1958 were identified from central population register data and surveyed in 1975 to form the Older Finnish Twin Cohort. Opposite sex pairs born 1938-1949 were included to expand the cohort in 1996. Follow-up data collection took place in 1981, 1990 and 2011-2012 [154, 156]. Several sub-studies concentrating on specific phenotypes such as nicotine dependence, alcohol addiction, and hypertension were more intensively studied (biological and clinical samples were collected from the participants) and are described below:

    NAG-FIN (Study I, III). Twins concordant for current smoking based on the questionnaire data were selected and recruited along with their family members (mostly siblings) to establish a family study designed to address genetics of nicotine dependence as part of an international consortium - Nicotine Addiction Genetics (NAG-FIN) [114, 157]. Data was collected in 2001-2005 and included a self-report questionnaire, diagnostic telephonic interview, and blood sample for DNA extraction. Information on lifetime smoking behavior aspects (including initiation, quantity, cessation, withdrawal symptoms, alcohol use and psychiatric comorbidities) were collected to extensively capture the complex landscape of addiction behaviors.

    Schizophrenia twin cohort (Study III). The SCZ twin cohort [158, 159] comprises of same-sex (both monozygotic and dizygotic) twin pairs discordant for schizophrenia diagnosis based on the Diagnostic and Statistical Manual of Mental Disorders, 4th Edition (DSM-IV)

criteria [160]. From this sub-study, 18 schizophrenia cases, and 55 controls (18 co-twins, and 37 control twins) had blood samples drawn and corresponding gene expression data available (N = 73).

EH-Epi (Study II). To study hypertension, individuals with diagnosed hypertension, use of anti-hypertensive medications and a history of blood pressure measurements were selected. These individuals were interviewed, examined and subjected to blood draws in 2012-2014. Methylation data from individuals (N=55) within this sub-study was available and utilized in study II.

2) **FinnTwin16 (Study I & II).** Initiated in 1991, FinnTwin16 is a longitudinal study comprising twins born during 1975-1979, first assessed at age 16 along with their parents. Follow-up data collection was done at ages 17, 18.5 and 24 [157] with a fifth wave of follow-up done in 2011 [154]. This study in younger twins was aimed at examining behavioral risk factors at an early age and later disease outcome. Along with self-report questionnaires and neuropsychological tests, blood samples for DNA extraction and serum metabolite measurements were also collected. For a subset of these twins who participated in sub-studies focusing on alcohol (Alcohol study, N=54) and obesity (TwinFat, N=14) research, methylation data was produced during the 4th wave of data collection and these samples were used in study II.

   Alcohol study. This sub-study within the FinnTwin16 was designed to study the effects of alcohol on brain [161]. Twin pairs concordant and discordant for alcohol use were selected based on the self-report questionnaire data from 2000-2002 follow-ups.

   TwinFat. This sub-study was based on individuals from FinnTwin12 and FinnTwin16 (see below) which were selected to study obesity in twins. These twins completed an intensive metabolic study protocol [162] and assessment of behavioral traits (questionnaires, interviews, and diaries) in 2002-2013.

3) **FinnTwin12 (Study I & II).** Initiated in 1994, FinnTwin12 is a longitudinal study comprising twins born during 1983-1987, first assessed at age 11–12 along with their parents and teachers to study the precursors of health-related behaviors, especially use and abuse of alcohol [163]. Follow-up data collection were done at ages 14, 17.5 years and 22 [157, 163]. Data collection included structured psychiatric interviews and blood for DNA extraction. Methylation and serum metabolite measurements for N=189 individuals from FinnTwin12 were available and included in the analysis for study II.

Table 4. *Samples included in the three studies and their characteristics.*

Study	Cohort	Age (Mean, Range, SD)	% Males	BMI (Mean, Range, SD)	Genotype	Methylation	Expression
I	FTC	24 (21-30, 2.0)	50	24 (16-43, 4.0)	385	171	-
	FINRISK2007	49 (25-74, 12.8)	54	27 (16-50, 5.0)	419	-	-
	YFS	30 (15-45, 8.6)	54	24 (16-46, 3.9)	714	-	-
II	FTC	30 (21-68, 14.2)	48	24 (16-42, 4.4)	304	310	-
III	NAG-FIN	56 (30-92, 10.1)	52	-	1998	-	-
	DILGOM	52 (25-74, 13.7)	46	27 (16-47, 4.7)	512	512	512
	SCZ	45 (24-57, 8.3)	30	-	-	-	73

SD: Standard deviation; BMI: Body Mass Index

In addition to the FTC, additional cohorts from Finland were included in study I for GWAS meta-analysis and in study III for functional annotation:

The Young Finns Study (Study I). With first data collection in 1980, the Young Finn Study (YFS) is a follow-up study of cardiovascular risk factors from childhood to adulthood [164]. The follow-up was done at 3, 6, 9, 12, 21, 27, and 30 years, with wide-ranging risk factor assessments, including smoking status and alcohol use. A total of 714 current smokers were included in study I for meta-analyses having both genotype and metabolite measurements available from 2001 follow-up when the average age of participants was 21 years.

FINRISK (Study I). Finnish adult population survey-based studies conducted every five years since 1972 assessing risk factors of chronic diseases [165]. At the end of 2006, adults (age 25-74) who had ever smoked, were given a smoking-specific questionnaire as a part of FINRISK survey in 2007. A total of 419 biochemically (cotinine level>10ng/ml) verified current smokers were included in study I meta-analyses. A larger non-overlapping sample (N=19,857) from the FINRISK cohorts (1992, 1997, 2002 and 2007) was additionally utilized in examining the linkage disequilibrium (LD) patterns and allele frequencies, and for prediction analysis of the GRS in study I. Metabolite values were not available in this larger sample.

DILGOM (Study III). DIetary, Lifestyle and Genetic determinants of Obesity and Metabolic syndrome (DILGOM) is a sub-study within FINRISK2007. It is a cross-sectional population

study designed to examine the effects of lifestyle factors (such as diet and exercise), environment, and genetics on obesity and metabolic syndrome [166, 167]. Data collection was done in 2007 which included self-report questionnaires as well as blood samples. For N=512 individuals, genome-wide genotype, gene expression and DNA methylation data generated at the same time point were available, making it a valuable multi-omics dataset.

Ethical permissions for all the cohorts have been approved by appropriate ethics committees and have been submitted to the Faculty of Medicine, University of Helsinki in conjunction with this thesis.

4.2 Phenotypes

The aim of the thesis was to dissect the genetics and epigenetics of smoking behavior. Across the three studies, a wide range of phenotypes capturing smoking behavior and associated comorbidities were tested. Table 5 provides a summary of the phenotypes tested segregated into three wide categories:

Biomarkers/Metabolites. The metabolites - cotinine and *trans* 3-hydroxycotinine are the metabolized products of nicotine. These were measured from the serum using liquid chromatography-tandem mass spectrometry at the University of Toronto, Canada (Prof. Rachel Tyndale's lab) and at the National Institute for Health and Welfare, Helsinki, Finland [168]. For a subset of the sample used in study III (N=68), cotinine measurements were done at the Metabolomics unit at Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Finland [169, 170]. The ratio of *trans* 3-hydroxycotinine to cotinine (NMR) was used in study I, whereas cotinine (nanogram per milliliter; ng/ml) was used in study II. The NMR values were rank transformed to achieve near-normal distribution as it was used as the dependent variable in the analyses. Rank transformation was performed using function 'rntransform' in R package GenABEL [171] which sets the median value as zero and transforms all values centering the normal distribution around it. In study I, a threshold of cotinine ≥ 10 ng/ml was applied to select current regular smokers, whereas in study II a lower threshold of cotinine ≥ 4.85 ng/ml was applied. The lower threshold of cotinine has previously been suggested as an appropriate cut-off to identify current smokers in Caucasians [172]. This was used to maximize the sample size (methylation data available) for analysis. To ensure using a lower threshold did not affect the results heavily,

we performed sensitivity analysis using the threshold of cotinine>10ng/ml (N=293) and observed negligible differences (Appendix I).

Self-reported smoking. Questionnaire-based data for the self-reported smoking and alcohol use was available from the NAG-FIN cohort. Detailed self-reported smoking status was available from DILGOM cohort [173] categorized as current daily smokers (N=84), current occasional smokers (N=34), recent quitters (1-6 months of abstinence) (N=13), former smokers (>6 months of abstinence) (N=133) and never smokers (N=245). NAG-FIN and DILGOM cohorts are valuable resources, particularly for the work included in this thesis, as they provide a comprehensive view of smoking behavior. The smoking phenotypes analyzed were adapted from Broms et al. [174] and are listed in Table 5. In the statistical analysis, cigarettes per day (CPD) was used as class means of CPD (1.5, 3.5, 8, 13, 17.5, 22.5, 32.5, and 45 CPD, respectively) such that the regression coefficients can be interpreted as the average change in number of cigarettes smoked per day when the number of minor alleles is increased by one. Even though quantitative phenotypes provide higher statistical power than binary phenotypes [175], in study III, the quantitative variables were converted to binary variables (see Table 5 for definitions) to accommodate the data for joint linkage and LD analysis and rare variant association analysis (see section 4.4.1.1). The cut-off of 3 or 4 was based on the number of criteria needed to make a diagnosis as specified in the DSM-IV manual and the manual for the semi-structured assessment for the genetics of alcoholism [176] and nicotine addiction genetics [177] interviews.

Smoking associated comorbidities. In study III, to examine the involvement of NSP specifically in smoking, we utilized the self-reported data on alcohol use and abuse in the NAG-FIN sample. Since SCZ and smoking are also highly comorbid, we utilized the SCZ twin sample to examine the confounding effect of smoking on the NSP gene expression between SCZ and controls.

Covariates. In all studies, relevant covariates were included in the regression models. Age and sex were always included in all analyses. For methylation data, we also included body mass index (BMI), and white blood cell type proportions inferred from the methylation levels of the samples using the houseman algorithm [178]. Sample descriptive (age and sex) are listed in Table 4.

Table 5. *A complete list of phenotypes tested in the three studies along with their definitions.*

STUDY	PHENOTYPE	SAMPLE SIZE (N)	MEAN (RANGE, SD) / CASES (%) [a]	DEFINITION
Smoking biomarker - Nicotine metabolism and exposure				
/	Nicotine metabolite ratio (NMR)	1518	0.4 (0.01-2.0, 0.23)	Proxy for the rate of nicotine metabolism; calculated as the ratio of 3-hydroxycotinine to cotinine.
//	Cotinine	310	192.7 (5.1-820.5, 148.43)	Reliable indicator of nicotine exposure. Serum cotinine levels (ng/ml) measured using mass spectrometry.
Self-reported smoking				
///	Smoking initiation	1998	1660 (83%)	Smoked at least 100 cigarettes in lifetime and at least once a week for at least two consecutive months.
///	Cigarettes per day (CPD)	1998	18.8 (1.5-45, 10.2)	Number of cigarettes smoked per day during the month of heaviest smoking; eight categories: 1-2, 3-5, 6-10, 11-15, 16-19, 20-25, 26-39, and ≥40 CPD. 2
///	Maximum CPD	1998	29 (0-98, 14.3)	Maximum number of cigarettes ever smoked during a day (24h period).
///	FTND (≥4)	1998	798 (40%)	Nicotine dependent if ≥4 out of 10 points in Fagerström Test for Nicotine Dependence.
///	FTND score	1998	3.5 (0-10, 2.4)	Fagerström Test for Nicotine Dependence score (range 0-10).
///	DSM-IV nicotine dependence diagnosis	1998	844 (42%)	DSM-IV diagnosis for nicotine dependence (≥3 symptoms out of 7 occurring within a year).
///	DSM-IV nicotine dependence symptom count	1998	2.9 (0-7, 1.7)	Number of DSM-IV nicotine dependence symptoms (range 0-7).
///	DSM-IV nicotine withdrawal diagnosis	1998	522 (26%)	DSM-IV diagnosis for nicotine withdrawal (≥4 symptoms out of 8 occurring within a year).
///	DSM-IV nicotine withdrawal symptom count	1998	2.3 (0-8, 2.1)	Number of DSM-IV nicotine withdrawal symptoms (range 0-8).
Smoking associated comorbidities				
///	Regular drinker	1998	1183 (59 %)	Drinks at least one alcoholic drink at least once a week
///	Heavy drinker	1998	856 (43%)	Drinks at least five or more alcoholic drinks once a week.
///	Maximum drinks	1998	14.7 (1-72, 9.8)	Maximum number of alcoholic drinks ever consumed in one day (24-hour period).
///	DSM-IV alcohol dependence diagnosis	1998	103 (5%)	DSM-IV diagnosis for alcohol dependence (≥3 symptoms out of 7 occurring within a year).
///	DSM-IV alcohol dependence symptom count	1998	1 (0-7, 1.6)	Number of DSM-IV diagnosis for alcohol dependence (range 0-7).
///	Schizophrenia (SCZ)	73	18 (25%)	Diagnosis of schizophrenia was assigned per DSM-IV criteria

[a] Mean, range and standard deviation are presented for quantitative traits while number of cases and % are presented for binary phenotypes.

4.3 Omics data

This section describes the different data layers analyzed in the three studies. All omics datasets were produced with microarray-based technology. Figure 3 on page 7, provides an overview of the workflow of the microarray. Methods for analyzing these data are described in the next section (4.4. Analyses). All the data was available at different stages of processing: Genotype data for all studies was available post-QC (described in section 4.3.1 below) and imputation for analysis. Methylation data used in Study I and III were available post-QC and normalization (detailed in section 4.3.2), whereas for study II data was processed from raw intensity files (detailed in section 4.3.2). Expression data for SCZ dataset were processed and analyzed at Yale University while expression data for study III (DILGOM) was preprocessed and normalized at Estonian Genome Center, University of Tartu and was made available for analysis. In the following sections, quality control and data processing for genotype, methylation, and gene expression data is described.

4.3.1 Genotype data

Genotype data (SNP data) used in this thesis, was produced at the Wellcome Trust Sanger Institute, UK and the Broad Institute of MIT and Harvard, USA. Figure 5 details the genotype data processing as well as standard quality control (QC) criteria applied to the genotyped and imputed

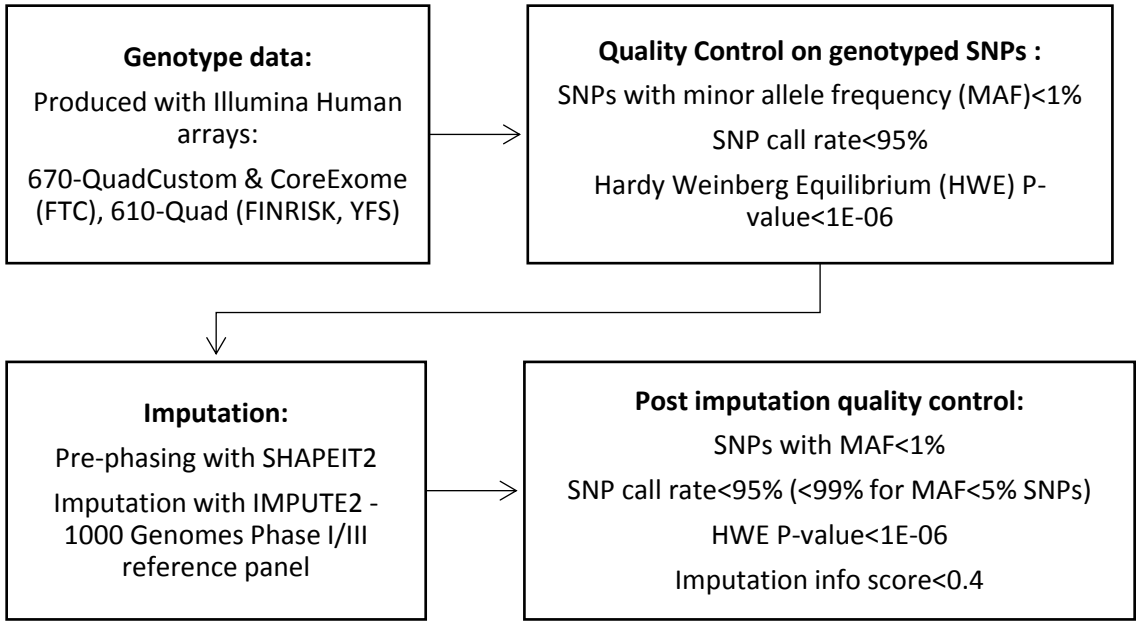


Figure 5. *Figure illustrating the process of genotype data processing including imputation to reference panel along with the quality control (exclusion) criteria pre-and post-imputation.*

data. QC and imputation for all genotype data were done centrally at the Institute for Molecular Medicine Finland (FIMM), University of Helsinki, Helsinki, Finland. In addition to post-genotyping-QC thresholds (detailed in Figure 5) applied to exclude bad quality SNPs, samples were excluded if call rate<95%, and if they failed in heterozygosity test and sex checks, or were outliers in Multidimensional Scaling plots. The QC thresholds in Figure 5 apply to all the arrays except for CoreExome chip where minor allele frequency (MAF) cut-off of minor allele count <2 was applied.

In study III, we tested the association between rare variants (MAF<0.01) and smoking as well as alcohol-related phenotypes. Given the low frequency of rare variants (power loss) and owing to the burden of multiple testing [179], we only analyzed variants located in coding and regulatory regions (enriched with phenotype associated rare variants [180]). Low frequency (MAF<0.05) and rare variants (MAF<0.01) were imputed, hence post-imputation QC criteria (Figure 5) were not applied to the NAG-FIN sample used in Study III. As study III is focused on the ten key genes in the NSP (Table 2 on page 12); genotypes were extracted based on the longest isoform reported at the UCSC Genome browser (per GRCh37/hg19) for each gene and 50kb flanking region. In all genetic association studies where FTC samples were used, only one individual from a monozygotic twin pair was included as they have identical genotypes, whereas both individuals from dizygotic pairs were included in the analyses.

4.3.2 Methylation data

Methylation data was generated using the Illumina Infinium HumanMethylation450 BeadChip (450k) at the Technology Centre, FIMM, Helsinki, Finland, the Microarray consortium, Oslo, Norway, The Genomics facility, University of Chicago, IL, USA and at the SNP&SEQ Technology Platform, University of Uppsala, Sweden. All samples were derived from peripheral blood, bisulfite converted using EZ-96 DNA Methylation-Gold Kit (Zymo Research, Irvine, CA, USA) according to the manufacturer's instructions and assayed on the 450k array. Figure 6 illustrates the processing of methylation data for analysis. To ensure good quality of data several QC metrics were applied. Probes with detection P-value>0.05 were discarded as per Illumina's recommendation to ensure captured signal was not background noise. Sample and probes were filtered out with a threshold of call rates>95%. For study II, detection P-value threshold was lowered to P-value≥10E-16 (such that probes on Y-chromosome were no longer detected among females) based on the recommendation by Lehne et al. [54]. Samples with gender mismatch (inferred based on the deviation from median intensities of probes mapping to X and Y

chromosome) were also excluded. To avoid spurious associations, probes that may be unreliable because of non-specific binding [181], presence of SNPs in probe body or at the CpG site [182], or insertions, deletions, and repetitive DNA [183] (listed as *ad hoc* exclusion criteria in Figure 6) were further excluded. Methylation data was used as beta value (percentage methylation) in all analyses. Calculated as the ratio of intensities between methylated *versus* combined locus intensity, and ranges between 0 (fully unmethylated) and 1 (fully methylated):

$$\textit{Methylation beta} = \frac{\textit{Methylated allele intensity (M)}}{\textit{Unmethylated allele intensity (U)} + \textit{Methylated allele intensity (M)} + 100}$$

Methylation data utilized in the three studies was pre-processed and normalized using the Bioconductor R packages 'minfi' [53] (study I, II and III) and 'limma'[184] (study II) with varying normalization methods and QC criteria as described below for each study. For sample QC based on number of bead counts 'wateRmelon' R package [185] was used. For annotation of the probes on 450k array, we used R package 'IlluminaHumanMethylation450kanno.ilmn12.hg19' and the data available by Zhou et al. [186].

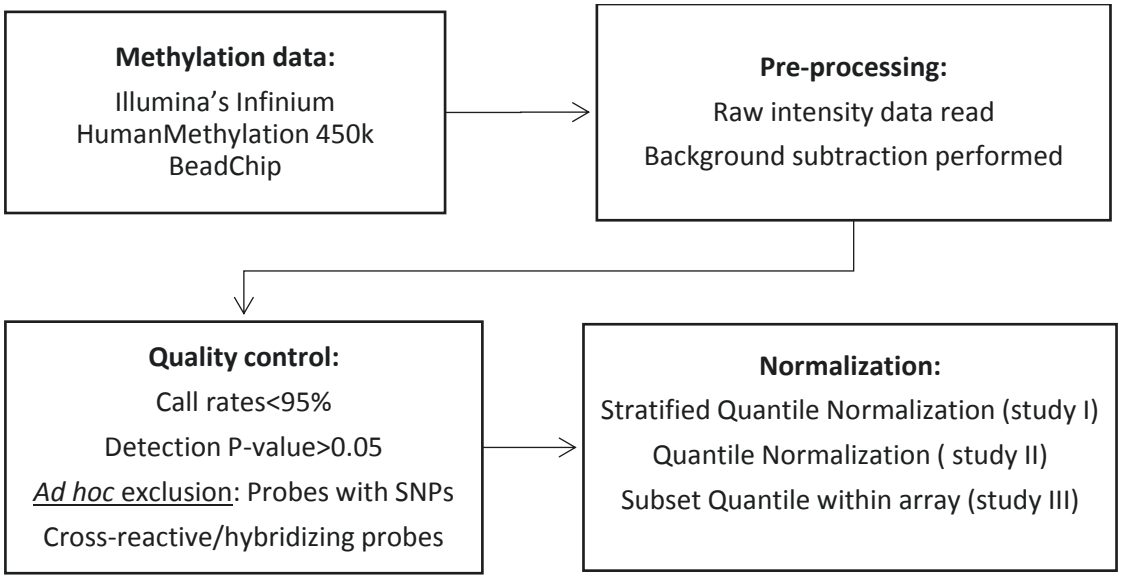


Figure 6. *Figure illustrating the methylation data pre-processing and normalization.*

In study I and III, only targeted regions were analyzed for assessing the role of methylation. In study I, CpG sites were chosen based on genomic coordinates mapping to the 4.2Mb (plus 500kb flanking) region containing the genome-wide significant loci identified in the GWAS meta-analyses of NMR. CpG sites mapping to this region were extracted from a 450k dataset processed with functional normalization [55] implemented in R package 'minfi' with the function 'FunNorm'

that uses control probes to remove unwanted technical variation as well as diminishes batch effects. A total of 1424 (out of 2268) CpG sites were taken forward for analysis after excluding unreliable probes. As the sample size for methylation data was small (N=171), further filtering was performed based on interquartile range (IQR) to only keep probes showing reasonable interindividual variability (IQR>0.05), leaving 158 sites for the analysis. In study III, CpG sites mapping to the ten NSP genes were selected from a Subset-Within Array normalized [52] dataset (DILGOM, N=512). The normalization was performed using 'preprocessSwan' function in 'minfi' R package which performs quantile normalization on the data based on subsets of probe types (I and II) as well as adjusts for the number of CpGs in the probe body. Altogether, 254 CpG sites mapped to the ten NSP genes (Table 2 on page 12), and 226 of these passed the QC criteria. In study II, the whole epigenome captured by the 450k array was analyzed using the pipeline suggested by Lehne et al. [54] with one additional QC criteria i.e. bead count>3 per sample. Post QC, N=310 samples and 323,478 probes remained. The intensity values were quantile normalized using the function 'normalizeQuantiles' in R package 'limma' [184] based on six different probe-type categories defined by color channel, probe-type and M/U subtype [54]. Normalized intensity values were used to calculate the methylation beta values at each CpG site which were used for the epigenome-wide analysis.

4.3.3 Expression data

Expression data from peripheral blood was produced with Illumina HT-12 expression array for the DILGOM sample (N=512) at Estonian Genome center, Tartu, Estonia and Illumina Human WG6 v3.0 Expression BeadChip for the SCZ sample (N=71) at University of California Los Angeles, USA. Expression datasets were used in study III for eQTL analysis and differential expression analysis of the ten NSP genes (Table 2 on page 12). Expression arrays were preprocessed with Illumina's BeadStudio software, followed by quantile normalization using R package 'Affy' [187]. Technical replicates were averaged to obtain one value per sample and log-transformed values were used for the analysis [166]. Probes were mapped to gene names using R package 'illuminaHumanv4.db'. Altogether, 19 probes mapped to the ten NSP genes, but only 17 passed QC criteria (probes mapping to non-autosomal chromosome, erythrocyte globin components or multiple genomic loci were removed) resulting in the exclusion of two probes mapping to *PSENEN* gene. One probe per gene was selected if it had the highest IQR for that gene to only test association with probes with most biological variation. Data from the SCZ sample was otherwise identically processed, except that rank normalization was applied and low signal strength probes

were excluded [188]. The same probes as selected in the DILGOM sample were picked from the SCZ sample.

4.4 Analyses

Analyses performed in the three studies are covered in this section. As summarized in Figure 1, the hypothesis-free approach was applied in study I and II, to examine the association of genome-wide genetic and epigenetic variants with NMR and serum cotinine levels, respectively. For study III, a hypothesis-driven, targeted approach was applied to extensively scrutinize association between genetic variants in the ten NSP genes (Table 2 on page 12) and a wide-range of smoking and alcohol use phenotypes (Table 5 on page 22). Integrative omics analyses to assess the potential function of variants associated with the phenotypes are also described in this section. Except for study I (GWAS-meta analysis of NMR), where a standard genome-wide significance threshold (P-value<5E-08) [189] was applied, statistical significance was declared when false discovery rate (FDR)[190] adjusted P-value<0.05, unless stated otherwise.

4.4.1 Association analysis (Study I, II & III)

With the aim to scrutinize the association of genetic and epigenetic variants with smoking behavior phenotypes we employed the following statistical methods:

4.4.1.1 Association of genetic variants

Association of common and low-frequency variants (Study I and III)

To assess the association of genetic variants (SNPs) with the quantitative traits, we employed univariate linear mixed model implemented in GEMMA (genome-wide efficient mixed-model association)[191] with the phenotype as the dependent variable and genotype along with covariates age and sex as fixed effects of the model.

$$Phenotype \sim \underbrace{SNP + Age + Sex}_{\text{Fixed effects}} + \underbrace{Relatedness\ matrix}_{\text{Random effects}}$$

Population stratification and genetic correlation (relatedness) in the sample was modeled with additional random effects using a covariance matrix estimated from the correlation (relatedness) of genome-wide genotype data across samples as implemented in GEMMA [191]. Only common (MAF>0.05) and low-frequency variants (0.05≥MAF≥0.01) were analyzed with GEMMA, and tests for rare variants (MAF<0.01) are described below.

Association of rare genetic variants (Study III)

In study III, we analyzed the rare variants (MAF<0.01) residing in coding regions, splice sites, promoters and untranslated regions of the ten NSP genes (Table 2 on page 12) with the aim to identify additional genetic contribution to the variation observed in the phenotypes tested. We performed single-variant and gene-based tests (as single-variant tests suffer from statistical power loss [192]). In single variant association analysis, we used 'lmekin' function implemented in R package 'coxme' [193] for quantitative traits and 'pedigreemm' [194] for binary traits to assess the association between phenotype and SNP with linear mixed effects model. To account for the relatedness in our sample a kinship matrix was included in random effects, while age and sex were added as covariates (fixed effects). Gene-based tests were performed using R packages 'SKAT' (SNPset (Sequence) Kernel Association Test) [195] for quantitative traits and 'HBMR' (Hierarchical Bayesian Multiple Regression model) [196] for binary traits. SKAT performs multiple regression to compute P-values based on the variance component of the aggregated set of SNPs in a region. HBMR performs multiple regression utilizing information on the relative contribution of variants toward the variance observed in the phenotype while also accounting for the genotyping quality. HBMR reports Bayes factors (BFs), and we considered nominal significance when BF>2.45, corresponding to a P-value<0.05 [197, 198].

Meta-analysis (Study I)

We performed a meta-analysis to combine the results from the three Finnish cohorts using META software employing the fixed effects model [99]. We used the 'inverse variance' method, wherein the effect size was estimated as the sum of beta estimates from each cohort weighted by the inverse of their sample variance and genomic control inflation factor.

Conditional analysis (Study I)

We performed conditional analyses to ascertain independent loci in the regions genome-wide significant signal in the GWAS meta-analyses of NMR. SNP with the lowest p-value in the meta-analysis was assumed to be the first independent loci. The genotype for this top SNP was then added as a covariate to the regression model in the separate cohorts and then meta-analyzed. SNP with the lowest P-value identified in this round was declared as the second independent signal. The top two SNPs were then together added to the regression model to further identify significant hits. This process was repeated until no further genome-wide significant association was observed.

Percentage of variance explained (Study I & II)

To estimate the proportion of variance explained by individual SNPs, we calculated the difference of R^2 between the two models i) Phenotype regressed only on the covariates (age, sex, BMI) and ii) model including the SNP along with the covariates.

Genetic Risk Score (Study I and II)

We calculated a genetic risk score (GRS) using the independent top SNPs from the meta-analysis in study I (rs56113850, rs113288603, esv2663194, and rs12461964; Table 7). The GRS was calculated as a weighted (by their estimated effect sizes) average of the major allele counts. The GRS was used to predict smoking behavior in two non-overlapping Finnish cohorts using logistic regression (to model current daily *versus* former smoking) and hurdle regression (to model smoking quantity among current smokers). In study II, we used this GRS to account for the effects of NMR on cotinine levels. For the sample included in study III, esv2663194 was not available, thus the GRS was constructed using only three SNPs (rs56113850, rs113288603, and rs12461964).

Joint linkage and LD analysis of common variants (Study III)

Linkage analysis and allelic association (also referred to as LD) when combined (joint linkage and LD) can be more informative in genome mapping. To utilize the extended family structure in our discovery sample (NAG-FIN) for study III, we scrutinized the genetic variants for linkage as well as joint linkage and LD with the PSEUDOMARKER software [199, 200] which implements the Elston–Stewart algorithm for full-likelihood. Only binary phenotypes and common variants (MAF≥0.05) were analyzed with PSEUDOMARKER assuming a recessive mode of inheritance.

4.4.1.2 Association of epigenetic variants (Study II)

To assess the association between methylation and serum cotinine levels, we performed epigenome-wide association analysis using pipeline suggested by Lehne et al. [54]. Modifications were made to the pipeline to accommodate the relatedness of our sample as explained below. In accordance with the pipeline, 30 principal components (PCs) based on the control probe intensities were used as covariates in the model to account for technical variance in the data. Intermediate residuals were estimated by regressing these 30 control probe PCs, age, sex and BMI on the methylation beta values. For any unaccounted global biological covariance, 10 PCs from based on the intermediate residuals were also included in the regression model. Houseman

algorithm [178] was employed to infer the white cell type distributions to be included in the regression model as well.

We employed linear mixed effect model using function 'lmer' in the R package 'lme4' [201] with methylation at each CpG as the dependent variable and serum cotinine level along with age, sex,

$$CpG \sim Cotinine + age + sex + BMI + white\ blood\ cell\ subtypes + 1\text{-}30\ PC_{control\ probes} + 1\text{-}10\ PC_{intermediate\ residual} + family + zygosity$$

Fixed effects — Random effects

BMI, white blood cell types and batch variable (for cotinine measurements performed at two different facilities; section 4.2) as covariates in the fixed effects part of the model. To account for relatedness in the sample, family and zygosity were included as additional random effects.

Secondary EWAS: Cotinine levels are affected by the amount of nicotine intake as well as the rate of nicotine clearance. To account for the influence of NMR on cotinine levels we utilized the NMR GRS (described above; Study I) as an additional covariate in the model.

Visualization of results (study I, II and III)

For Manhattan and QQ plots, R package 'qqman' [202] was utilized. LD structure of the SNPs highlighted in the analysis was visualized using Haploview software[203]. LocusTrack [204] was used to create regional plots for top associations. Circos plot depicting meQTLs were created using the R package 'Rcircos' [205].

4.4.2 Differential expression and methylation analysis (Study III)

As our association and joint linkage and LD analysis, revealed association between variants NSP genes and smoking behavior, we examined whether gene expression and methylation levels in these genes differ between smokers and non-smokers. Differential expression and methylation analysis were performed using linear regression ('lm' function in R) while adjusting for age, sex, and BMI. We further adjusted for estimated white blood cell proportions while comparing methylation levels. Owing to the high comorbidity between smoking and SCZ [80], we also examined the confounding effect of smoking on differential expression of the NSP genes in SCZ. We examined differential expression between SCZ cases and controls, adding age and sex as covariates while adjusting for relatedness in the sample with a kinship matrix. We then added smoking status (defined as smoker *versus* non-smoker) as a covariate to detect any confounding effects.

4.4.3 Quantitative Trait Loci analysis (Study I, II & III)

To assess the effect of SNPs associated located in the non-coding region of the genome, we performed methylation (meQTL) and expression quantitative trait loci (eQTL) analyses. In study I, we examined the association between SNPs and methylation levels of the CpG sites in the region with the genome-wide significant signal, by regressing methylation beta values against the genotype count of the coded allele (0, 1, or 2), while accounting for age and sex. In study II & III, we utilized the R package 'MatrixEQTL' [206] with the linear model setting to examine the association of the highlighted SNPs with methylation levels (meQTLs). Age, sex, BMI, smoking status (cotinine levels in study II) and white blood cell subtypes were included as covariates. In study II, we also included a covariance matrix based on the methylation levels of the top CpGs to account for relatedness in our sample. As opposed to study I and III, where the regions analyzed were targeted and we considered only *cis* interactions, in study II we also tested *trans* interactions. A *cis* distance of 2.5Mb (longest gene was ~2.4 Mb) was chosen to accommodate all possible combinations of SNP-CpG pairs for a given gene, while all other interactions were considered *trans.* In study III, we also assessed the association of top SNPs with expression levels of the NSP genes (eQTLs). We used identical procedure as for meQTLs except that white blood cell counts were not included in the model (to have comparable results for replication in publicly available datasets/resources).

4.4.4 Mediation analysis (Study I & II)

To investigate if methylation is a mediator between the observed association of genotype and phenotype, we performed causal inference test (CIT) [34] (Figure 7). CIT was conducted only for candidate SNPs that met the two basic conditions of association with the phenotype (NMR or cotinine levels) as well as methylation, the potential mediator. As illustrated in Figure 7, we examined the mediation by methylation between the observed association of SNPs with NMR (study I) and cotinine levels (study II).

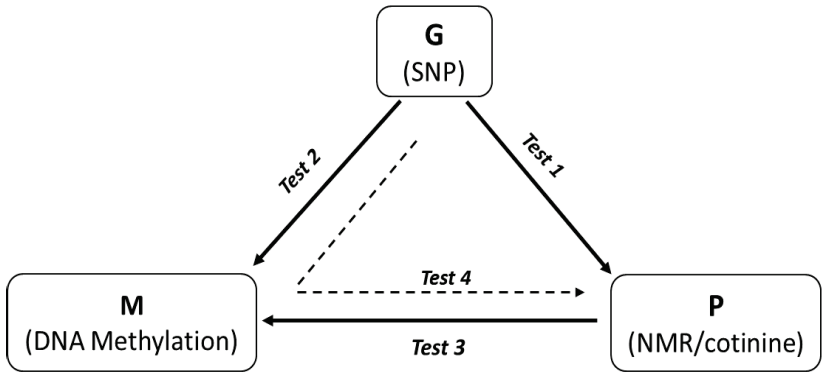


Figure 7. *Illustration of mediation analysis setup performed using causal inference test (CIT), a mathematical framework used to assess mediation. CIT tests conditional association for genotype (G), potential mediator (M, methylation), and phenotype (P): (i) G is associated with P, (ii) G is associated with M, (iii) P is associated with M conditional on G and (iv) P is independent of G conditional on M.*

4.4.5 Annotation of association findings (Study I, II and III)

In addition to employing in-house data, we also utilized several tools and databases (summarized in Table 6) for inferring predicted functional potential, pathway analysis, replication of our findings, and assessing tissue specificity of our results. All listed resources, except IPA, are public and freely available.

Table 6. *Tools and databases utilized in performing annotation of highlighted associations.*

Study	*Resource*	*Description*	*Weblink*
I, III	Variant effect predictor (VEP)	Provides information on the potential effect of the variants within genes as well as regulatory regions [207]	www.ensembl.org/Tools/VEP
II	Ingenuity pathway analysis (IPA)	Provides information on the enrichment of metabolic and signaling pathways in the data. [208]	www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis/
III	GTEx	Contains genotype and expression profiles across 53 different tissues (as of September 2017) enabling eQTL studies across tissues [209]	www.gtexportal.org
	BRAINEAC	Brain eQTL Almanac provides genotype and expression profile across 10 brain regions [210]	http://braineac.org/
	mQTLdb	Database with methylation and genotype data on mother-child pairs providing access to meQTL mapping across five different stages of life in two tissues (peripheral and cord blood) [211]	www.mqtldb.org/
	Fetal brain meQTLs	Epigenome-wide significant meQTLs observed in fetal brain [212]	http://epigenetics.essex.ac.uk/mQTL/
	Schizophrenia Genetic Research database (SZDB)	Collective database for SCZ research containing genetic, gene expression, network-based data from several studies. [213]	www.szdb.org
	HaploReg v4	Provides regulatory potential of non-coding SNPs [214]	www.encodeproject.org/software/haploreg/
	SPANR	Splicing disruption potential of coding and non-coding variants [215]	http://tools.genes.toronto.edu/
	450k Annotation	Comprehensive annotation data for 450k array [186]	http://zwdzwd.github.io/InfiniumAnnotation

5 Results & Discussion

5.1 Study I: Genetic variants associated with nicotine metabolism rate and their interplay with methylation

GWAS using metabolites measured from serum such as glucose, insulin, and lipids have been highly successful in identifying underlying genetic risk factors [72, 73] and therefore represent powerful intermediate phenotypes for complex traits. To identify novel genetic variants associated with nicotine metabolism rate, using a ratio of nicotine metabolites 3-hydroxycotinine and cotinine (NMR), we performed a GWAS meta-analysis with three Finnish cohorts (N=1518). We observed a strong association signal at 19q3.2 locus. A total of 719 SNPs reached genome-wide significance (P-value<5E-08) within a 4.2 Mb region (hg19 build coordinates 19:39546965–43710562). Manhattan plot for the NMR GWAS meta-analysis is presented in Figure 8.

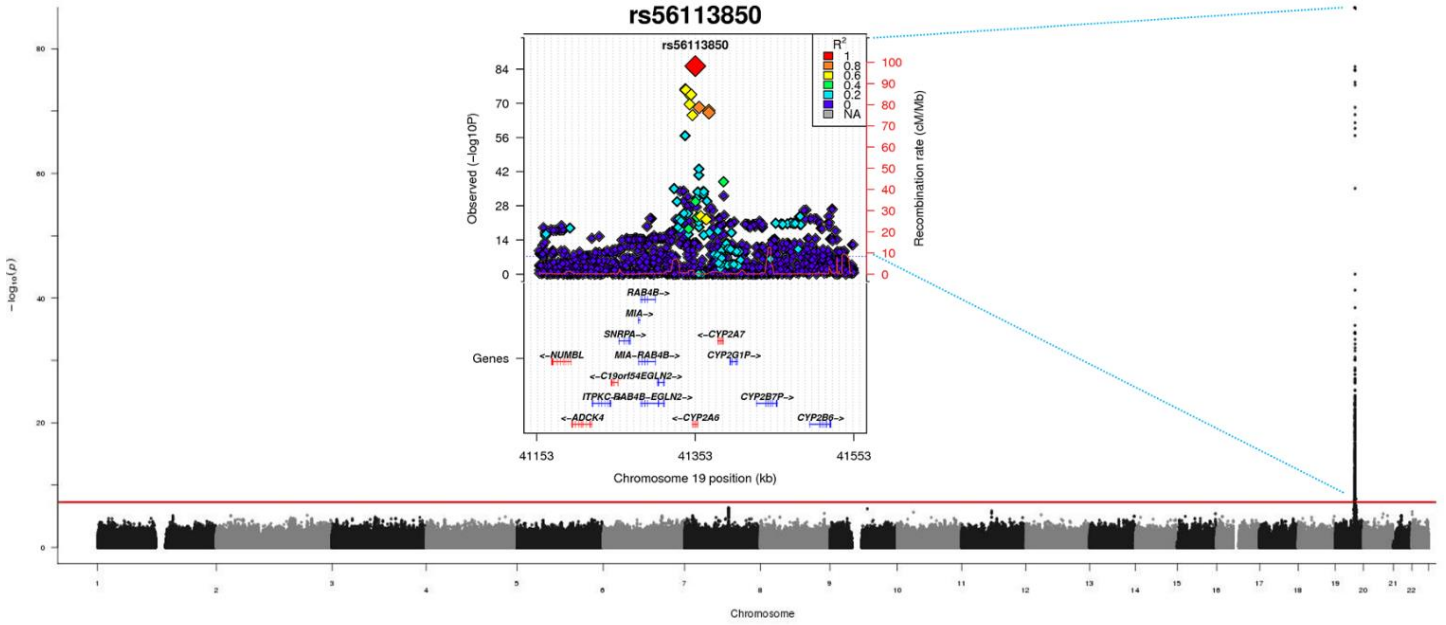


Figure 8. *Manhattan plot for the NMR GWAS meta-analysis with zoomed in region on 19q3.2. Figure modified from Loukola et al. PloS Genet 2015.*

A SNP in the intron 4 of *CYP2A6,* rs56113850 (referred to as index SNP here on), was the top association in the meta-analysis (P-value=5.77E-86). Conditional analysis of this region revealed three independent SNPs (Table 7), with rs12461964 appearing as a fourth independent signal only in the FINRISK2007 sample. Rs113288603 (not genome-wide significant in GWAS

meta-analysis) emerged as an independent signal only when the conditional analysis was performed on the index SNP. An interplay between the index SNP (rs56113850) and rs113288603 was observed, such that adding the index SNP to the model changed the association from non-significant (Beta=0.05, P-value=0.522) to genome-wide significant (Beta=-0.47, P-value=1.32E-09). The index SNP also showed an increase in the magnitude of effect size from -0.65 to -0.82 (P-value<2E-16) when rs113288603 was added to the model. Adding an interaction term between rs113288603 and the index SNP further increased the magnitude of effect size for rs113288603 (Beta=-0.62, P-value=1.32E-03) but not for the index SNP (Beta=-0.82, P-value <2E-16). These results indicate that the index SNP and rs113288603 act in concert and their interplay should be examined in future studies.

Table 7. *Independent signals identified with conditional analysis.*

SNP	*Position*	*Reference allele*	*Minor allele*	*MAF* [c]	*Beta (minor allele)*	*P-value* [d]
rs56113850	chr19:41353107	T	T	0.44	-0.65	5.77E-86
rs113288603 [a]	chr19:41362293	C	T	0.15	-0.02	7.03E-25
esv2663194	chr19:41355733	G	<DEL>	0.03	-1.08	8.49E-20
rs12461964 [b]	chr19:41341229	A	A	0.45	-0.61	3.66E-16

[a] *No association in the GWAS meta-analysis, but appeared as an independent signal in analyses conditioned on the index-SNP*

[b] *Emerged as an independent SNP only in the FINRISK2007 sample*

[c] *Minor Allele Frequency (MAF) calculated from a large FINRISK sample (N=19,857)*

[d] *P-values from each meta-analysis conditioned progressively in the order of SNPs listed in the table.*

All the independent SNPs were in or near (≤8kb) *CYP2A6*, the main nicotine metabolizing enzyme. Minor alleles for all top variants had a decreasing effect on NMR corresponding to a decrease in nicotine clearance rate (Table 7). The index SNP (rs56113850) by itself, explains a fairly large proportion of variance in NMR ranging between 14-25% in the three cohorts included in the study and has a prominent effect size (Beta=-0.65). Similarly, the indel esv2663194 has a large effect size (Beta=-1.08). Together the three independent SNPs (rs56113850, rs113288603, and esv2663194) explain 20.8% variance in FinnTwin, 31.4% in YFS, and 26.3% in FINRISK2007 cohorts. In FINRISK2007, including the fourth independent SNP rs12461964 increased the variance explained to 27.7%.

We constructed a weighted genetic risk score (GRS) based on the independent SNPs and tested if it could predict smoking behavior in the NAG-FIN and non-overlapping FINRISK samples.

The GRS was calculated using the major allele counts and corresponding effect sizes, such that higher GRS parallels faster nicotine metabolism. First, we tested if the GRS associates with cigarettes per day (N = 3954) and second with smoking status i.e. current smokers (N= 3954) *versus* former smokers (abstinence of ≥ 6 months; N= 3543). Results were comparable for the GRS with three and four SNPs. The GRS was positively associated with cigarettes per day in current smokers (Beta=0.10, P-value=0.002), indicating that individuals with faster metabolism smoke more, in line with earlier evidence [216]. The GRS was also associated with increased likelihood of being a former smoker (OR=1.39, 95% CI 1.09 −1.76, P-value=0.007). However, in the sensitivity analysis, when effects of potential confounders for smoking cessation (adverse health effects/diagnosis of major depressive disorder, cancer, or cardiovascular diseases) were taken into consideration, and the model was adjusted for alcohol consumption, the association became non-significant (OR=1.30, 95% CI 0.95-1.78, P-value=0.10). This may be due to lack of statistical power or GRS not capturing certain relevant aspects of smoking cessation success, preventing us from replicating previous findings that show slow metabolizers quit more often than normal metabolizers [105, 217]. However, the GRS still captures a considerable proportion (~30%) of the variance in NMR and can be of great utility in clinical settings to personalize cessation therapeutics based on genetically determined nicotine metabolism profiles [218].

For annotation of the probable role of the genome-wide significant SNPs, we considered the following hypotheses (results summarized in Figure 9): SNPs directly affect the function of the *CYP2A6* enzyme or are in LD with known functional variants of *CYP2A6*, or epigenetic mechanisms like methylation regulate the gene function. According to variant effect predictor (VEP; Table 6), the four independent SNPs did not have any predicted functional effects. Except for rs113288603, the other three independent SNPs shared LD blocks with the known functional variants of *CYP2A6* (https://www.pharmvar.org/htdocs/archive/cyp2a6.htm)[106]. To assess the role of epigenetics (methylation), we performed meQTL analysis and identified 173 SNPs associated with methylation levels at 16 CpG sites in the region harboring genome-wide significant association signal. We further performed CIT to examine whether methylation at any of these sites mediates the effect of genotype on NMR. We observed that methylation at the CpG site cg08551532 (in *DLL3* gene) mediates the effect of SNPs on NMR. Our meQTL and CIT results indicate that epigenetic mechanisms such as methylation may play a role in regulating at least some of the genes identified in this GWAS. It must be noted that the most interesting gene, in this case, *CYP2A6* could not be tested for meQTLs as none of the CpG probes mapping to it passed

QC. Targeted exploration of methylation in *CYP2A6,* for example using Epityper or bisulfite sequencing, would be required for further investigation of this region.

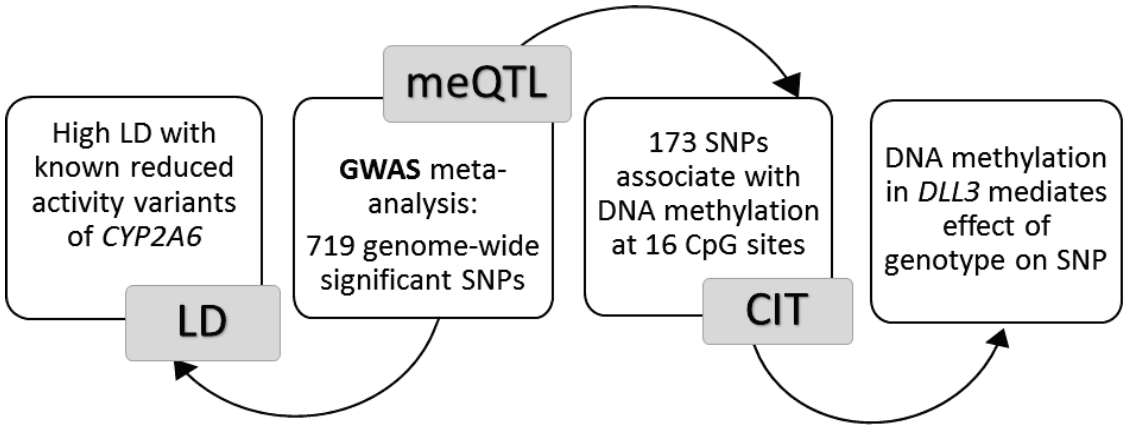


Figure 9. *Summary of annotation results for the genome-wide significant SNPs identified in GWAS meta-analysis.*

Although we examined epigenetic mechanisms potentially involved in mediating the effects of NMR associated SNPs, further investigation of these loci with extensive functional assessments would be valuable. For instance, other forms of genetic (rare variants, copy number variants, etc.) and epigenetic (histone modifications) variations, that were not captured in our study are essential. Sequencing of the highlighted locus for genetic and epigenetic variants could provide a better resolution into this region of the genome and identify additional factors contributing toward the variance in nicotine clearance rate and its consequent effects on smoking behavior and associated disease risk.

This study was performed in Finnish population samples, and the results may be specific to Finnish population. Replication of our findings in other populations is of importance. Two recent studies in European, African, and Asian American individuals by Baurley et al. [108] and in African Americans by Chenoweth et al. [109] also performed GWAS meta-analysis of NMR. Baurley et al. [108] replicated our top association, however, less than 40% overlap was observed with Chenoweth et al. [109] findings. Only one of the three independent SNPs in the African American study [109], rs12459249 was genome-wide significant in our meta-analysis (P-value=3.28E-76). These results provide support to our findings as well as highlight the population specificity of nicotine metabolism [103]. Observed differences may also be due to different genotyping platforms and imputation panels used in the three studies, and to a smaller extent because of hormonal (estrogen-induced *CYP2A6* activity) and demographic factors such as age, sex, BMI, alcohol, and cigarette consumption [219].

Our study was the first GWAS on NMR, where we identified robust associations at 19q3.2 locus encompassing the main nicotine metabolizing enzyme *CYP2A6* including several novel associations. Our study also demonstrates the value of a biomarker as a phenotype in a GWAS meta-analysis with modest sample sizes. Our findings of novel genetic variants associated with NMR and the possible role of epigenetic variation in the highlighted locus hold great value for aiding personalization of cessation therapeutics (as discussed in chapter 6).

5.2 Study II: DNA methylation associated with serum cotinine levels of regular smokers

As demonstrated by numerous EWAS (Table 3 on page 13), smoking has been extensively associated with DNA methylation levels. Self-reported smoking status and quantity have been widely utilized in such EWAS, identifying genes such as *AHRR*, *ALPPL2*, and *F2RL3* consistently. However, self-reported smoking status and quantity are prone to error due to misreporting and recall bias [66, 220]. Cotinine, on the other hand, is a reliable indicator of nicotine exposure [66]. Thus, we utilized serum cotinine levels as a continuous phenotype in an EWAS, performed in a sample of biochemically verified regular smokers (serum cotinine>4.85ng/ml), allowing us to assess the direct effects of nicotine exposure on methylation. We further evaluated the role of genetic variants in the genes identified in the EWAS and assessed whether methylation at the identified loci is a cause or consequence of nicotine exposure. Figure 10 summarizes the study design and the main findings.

In our EWAS of regular smokers, we identified methylation at 50 CpG sites (in or near 35 genes) significantly associated with cotinine levels. Among the 50 CpG sites, 33 CpG sites have previously been reported while 17 are novel (never reported as genome-wide significant in previous studies; Table 3 on page 13). Nine of the previously reported CpG sites were identified only in studies using self-reported smoking status with much larger samples (~1000 or more) [135, 141, 142], in contrast to our sample of 310 individuals, showing the power afforded by biomarker (a reliable and informative phenotype). An extended table of results from the EWAS is in Appendix I.

The top association was observed at cg05575921 in *AHRR* gene (P-value=3.3E-18), the most consistently reported association (Table 3). In line with earlier reports, we also observed CpG sites in *ALPPL2* (minimum P-value= 1.9E-15 for cg05951221) and cg03636183 in *F2RL3* (P-value= 1.9E-11) [133, 139, 141, 221]. Also, consistent with previous studies, for a substantial portion of the highlighted CpG sites (42 out of 50), methylation levels were negatively associated with cotinine levels; parallel to lower methylation with smoking exposure in other studies. A key observation is that the novel associations identified in our EWAS were in smoking-related genes such as cg13740236 in *LSM6* (P-value= 2.9E-08), cg05992400 in *CYP2C18* (P-value=4.9E-06), and cg26589665 in *THSD4* (P-value= 3.8E-07) covered in the following sections.

As cotinine level is influenced by both nicotine intake as well as nicotine clearance rate, we added a three SNP GRS, based on the findings from study I, to account for nicotine metabolism rate to the model. In this secondary EWAS, methylation at a total of 30 CpG sites (in or near 20 genes) was significantly associated with cotinine levels, five of which were novel (never reported as genome-wide significant in previous studies; Table 3 on page 13) and 25 overlapped with the discovery EWAS results. Interestingly, when the GRS was added, half of the associations from the discovery EWAS were no longer genome-wide significant (FDR P-value>0.05) in the secondary EWAS. This suggests that these genes are likely influenced by or are involved in nicotine metabolism. Pathway analysis of these genes (genes no longer genome-wide significant in the secondary EWAS), were enriched in xenobiotic and nicotine degradation metabolic pathways. The two genes mainly involved in these pathways are *CYP2C18* and *AHRR*, both related to the cytochrome P450 (CYP) family - a major source of variability in drug pharmacokinetics and response [222]. *AHRR* is a regulator of CYP genes. This show that the GRS accounts for the effect of nicotine metabolism on cotinine levels. However, the GRS based only on three SNPs does not capture the entire effect of NMR as a high proportion (~70%) of the variance in NMR remains unaccounted. A GRS composed of more SNPs (explaining more variance) or a more composite measure that could incorporate contribution of other mechanisms (such as epigenetics) that explain variance in NMR, would be ideal to account for the influence of nicotine metabolism rate more comprehensively.

A total of 55 unique CpG sites (in or near 40 genes) were identified in the EWAS (discovery and secondary) that were associated with cotinine levels. We observed that some of the highlighted genes in our EWAS harbor genetic variants previously reported for association with smoking-related phenotypes, for example, SNPs in *THSD4*, *LSM6,* and *CNTNAP2* are associated with nicotine dependence [223], and *LSM6* and *CACNA2D4* with pack years (an indicator lifelong accumulated smoking exposure) [224]. *THSD4* variants are also associated with smoking cessation success [225] as well as lung function affected by smoking [226]. Hence, we further explored the role of genetic variants in the highlighted genes. In our data, SNPs in *CNTNAP2* and *THSD4* genes were significantly associated with cotinine levels, in line with these previous findings. Altogether, 20 SNPs in nine genes were associated with cotinine levels, and strongest association was observed for rs187669467 in *ARHGAP44* (P-value=3.9E-06).

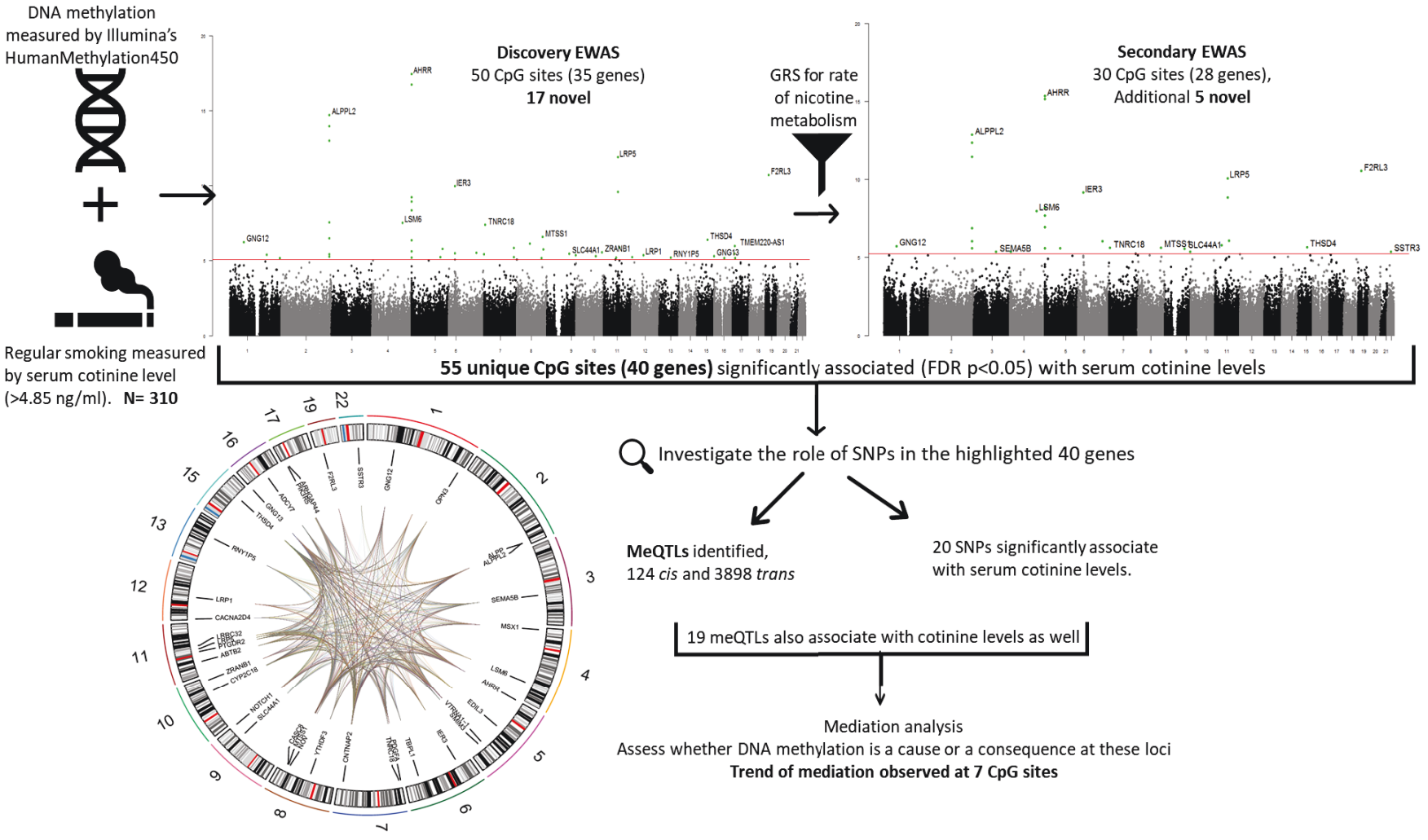


Figure 10. *Illustration of the study design and summary of results from the EWAS of serum cotinine levels in regular smokers*

In the 40 genes that were identified in our EWAS, we also examined the association of genetic variants with methylation levels of the top 55 CpG sites and identified 124 *cis* and 3898 *trans* meQTLs demonstrating both proximal and distal interaction between nicotine exposure associated methylation and SNPs. Interestingly, among the 20 SNPs that were associated with cotinine levels, 19 were also meQTLs. We performed CIT to investigate the role of DNA methylation as a mediator between the association of genotype and cotinine levels at these loci. We observed a trend (P-value<0.05) of mediation at 7 CpG sites with SNPs in distal genes, suggesting that methylation at these loci may not be a consequence of nicotine exposure but rather a causal mediator in the pathway between genotype and cotinine levels. These findings are crucial and require experimental validation.

We replicated several prior associations reported between smoking and methylation, the novel associations identified in our study point to exposure relevant loci, but require replication in other population samples. This study was conducted in a sample of regular smokers in contrast to earlier EWAS, including the two EWAS performed using cotinine [129, 140], that included both smokers and non-smokers. This major difference may be reflected in our novel findings, meaning that our novel findings might be specific to persistent/heavy smoking. A note of caution, as cotinine reflects recent nicotine intake (half-life 15-20 hours), cotinine values are similar (~zero) for never, former and occasional smokers while their methylation profiles may not be identical (as noted by several EWAS, listed in Table 3). Therefore, including individuals with cotinine levels ~zero should be avoided for methylation studies. Other sources of nicotine including snus, e-cigarettes, and nicotine replacement therapy may also contribute to cotinine levels making the results not entirely specific to tobacco smoking. However, the chances of such alternative sources of nicotine in our sample are low, since, at the time of sample collection, the use of such sources was rare in Finland. Other potential sources of discrepancies in the EWAS results may be attributable to the difference in sample size, population specificity of methylation levels [131], the precision of cotinine measurement technique (higher precision with mass spectrometry *versus* immunoassay) [227], or difference in methylation data preprocessing. Further, our modest sample size and use of 4.85 ng/ml threshold to define current smokers (in contrast to more commonly used 10ng/ml, like in study I) may have introduced some noise in the association results, however, since several of our novel associations point to smoking-related genes the likelihood is small. To further verify this, we performed sensitivity analysis using a sample of

smokers with cotinine>10ng/ml (N=293) and observed negligible differences in the beta estimates for all highlighted CpG sites (Appendix I).

In conclusion, our results showcase the power of a reliable and informative phenotype (biomarker) in identifying direct effects of nicotine exposure (smoking) on methylation levels. We replicated several earlier findings as well as identified novel biologically relevant loci, which could be potential targets for smoking cessation therapies. We also show that many of the identified methylation loci may be driven by underlying genetic variants and that methylation may be on the causal pathway for such smoking associated genetic variants rather than a consequence of nicotine exposure.

5.3 Study III: Smoking behavior associated functional variants in neuregulin signaling pathway

The NSP, involved in neuronal migration and differentiation, has been implicated in smoking behavior phenotypes (SI, ND, and NW) in both a behavioral mouse model and human studies [114-116] (detailed in section 2.5.1). We hypothesized that several risk variants in the ten key functional components of the NSP (Figure 4, Table 2) might be involved in smoking behavior. Given the high comorbidity between smoking and other addictions, for example, alcohol [77], we also examined if NSP is specifically involved in smoking and not co-occurring addictions. We tested association, linkage, and joint linkage & LD between common (MAF>0.05), low frequency (0.01≤MAF≤0.05) and rare (MAF<0.01) genetic variants in the 10 NSP genes (Table 2) with a wide range of smoking and alcohol use phenotypes (nine smoking-related phenotypes and five alcohol-related; Table 5 on page 22) in the NAG-FIN sample. Alcohol-related phenotypes were analyzed to assess the specificity of NSP toward smoking, wherein alcohol-use phenotypes were representative of other addictions. Altogether 66 SNPs in seven of the ten NSP genes residing in 23 distinct LD blocks showed significant association with three distinct phases of smoking behavior i.e. smoking initiation, nicotine dependence, and withdrawal. A summary of the associations identified and their functional annotation is provided in Figure 11. There was no association observed with any of the alcohol-related phenotypes, suggesting NSP's involvement in smoking behavior, ruling out any confounding effects of alcohol use. Testing other addictions would have been ideal to further support our hypothesis (NSP's involvement specifically in smoking), our study sample had limited use of other addictive substances like cannabis [228].

We observed that most of the association signal was confined to common variants and the SI phenotype, which is not surprising given that ~80% of our sample had initiated smoking providing higher power in comparison to fewer cases of ND (42%) and NW (26%) (Table 5). Only one SNP, rs13385826 in *ERBB4* was associated with ND (P-value=1.7E-07) in these analyses. Common and low-frequency variants in *NRG3* and *ERBB4* showed association with NW symptom count with large effect sizes (Beta range: -0.5 to -0.8). In single rare variant association analysis, no significant association was observed likely due to lack of power [179]. Gene-based rare variant association analysis indicated association between SI and variants in *NRG1* and *PSEN1*, and NW with variants in *ERBB4*, demonstrating the cumulative contribution of rare variants in these genes.

Gene	SNP	Chr	LD block	Effect allele	MAF	P-value **	Functional annotation
APH1A	rs183423866	1	1	G	0.064	1.8E-09	
	rs187135585	1		G	0.131	6.0E-06	
PSEN2	rs10916053	1	2	A	0.396	2.1E-04	
ERBB4	rs1836721	2	3	A	0.068	1.9E-05	
	rs13385826	2		A	0.128	8.0E-27	
	rs184443533	2	4	T	0.117	3.2E-09	
	rs6729616	2	5	A	0.057	4.9E-05	
	rs187745320	2		A	0.137	4.5E-07	
	rs13413099	2	6	A	0.216	4.0E-27	
	rs71350752	2		T	0.304	1.3E-07	
	rs192584214	2		T	0.065	2.4E-07	
	rs112465988	2		T	0.052	1.3E-05	
	rs185444317	2	7	T	0.081	1.0E-06	
	rs10167304	2		T	0.052	1.9E-04	
	rs12694263	2		C	0.050	2.1E-04	
	rs1581385	2		A	0.364	6.7E-09	
	rs187445356	2	8	T	0.053	6.1E-13	
	rs4993337	2	9	A	0.314	2.2E-07	
	rs141486719	2		T	0.182	1.1E-04	
	rs78391944	2		T	0.028	1.8E-05	
	rs71422776	2		A	0.035	3.5E-05	
	rs12990946	2		G	0.035	2.9E-05	
	rs35225704	2	10	C	0.033	9.7E-06	
	rs73989053	2		G	0.148	2.7E-06	
	rs13006797	2		G	0.136	2.1E-06	
	rs17328083	2		T	0.148	2.7E-06	
	rs13001305	2		T	0.126	2.2E-06	
	rs75489550	2		G	0.094	1.6E-05	
	rs36001505	2	11	C	0.061	1.9E-05	
	rs147786642	2		G	0.059	1.0E-05	
NRG1	rs113317778	8	12	G	0.096	1.3E-04	
	rs4327825	8	13	A	0.325	9.6E-10	
	rs4329235	8		C	0.284	9.3E-13	
	rs4436087	8	14	G	0.145	1.4E-04	
	rs10954844	8		A	0.152	1.7E-04	
	rs77085193	8		C	0.152	2.3E-04	
	rs17643260	8		T	0.152	2.3E-04	
	rs7000656	8		T	0.152	2.4E-04	
	rs17716671	8		A	0.152	2.4E-04	
	rs11780575	8		G	0.152	2.4E-04	
	rs17644185	8		A	0.152	2.4E-04	
	rs17716920	8		A	0.152	2.4E-04	
	rs10808325	8		T	0.152	2.4E-04	
	rs75974189	8		C	0.152	2.4E-04	
	rs11779465	8		G	0.145	2.2E-04	
	rs58182668	8		C	0.145	2.2E-04	
	rs11787140	8		C	0.152	1.9E-04	
	rs4733344	8		G	0.152	1.9E-04	
	rs12545304	8		A	0.152	2.4E-04	
	rs11777461	8		C	0.151	2.2E-04	
	rs75673683	8		G	0.152	2.4E-04	
	rs4733346	8		C	0.152	2.4E-04	
	rs7840005	8		T	0.153	2.7E-04	
	rs77752517	8		C	0.148	1.6E-04	
NRG3	rs79886978	10	15	T	0.081	9.8E-05	
	rs12774918	10		T	0.054	3.6E-07	
	rs11528179	10	16	A	0.074	2.9E-07	
	rs191742951	10	17	A	0.151	1.5E-05	
	rs11192578	10	18	A	0.038	1.2E-06	
	rs34434628	10	19	C	0.104	6.0E-06	
	rs60197472	10		T	0.390	5.6E-05	
	rs59013794	10	20	G	0.103	6.0E-06	
	rs7908968	10		G	0.166	2.9E-05	
BACE1	rs1261780	11	21	T	0.149	6.8E-05	
	rs191109295	11	22	G	0.191	4.0E-06	
PSENEN	rs807483	19	23	T	0.354	1.7E-04	

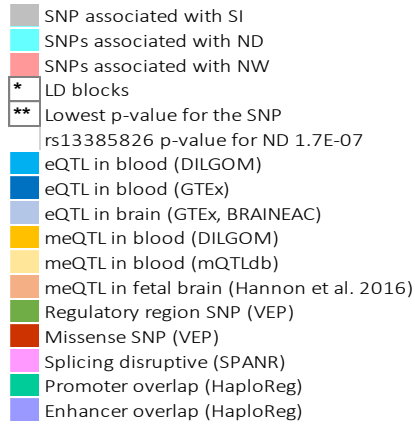


Figure 11. *Modified figure from Gupta et al. Transl Psychiatry 2017 showing the functional annotation of the 66 variants associated with smoking initiation, nicotine dependence, and nicotine withdrawal.*

Our results indicate that *ERBB4* gene is associated with different measures of smoking behavior (smoking initiation, nicotine dependence, and withdrawal), in line with previous findings [114, 115], as well as provide empirical evidence for further investigation of *ERBB4*'s role in nicotine addiction. Our results also show the involvement of *NRG3* in nicotine withdrawal symptoms; supporting the findings from a behavioral mouse model study [116] that also reported variants in *NRG3* were associated with successful smoking cessation in humans [116].

Except for a missense variant, rs113317778 in *NRG1*, all other variants identified in this study were in the non-coding region of the genome having unclear function or consequence. We performed comprehensive functional annotation of the 66 highlighted variants using epigenetic and transcriptomic data from an independent Finnish population sample (DILGOM), as well as publicly available data resources (Table 6). As evident from the visual representation of the results in Figure 11, we observed functional potential in a large proportion (56/60) of the highlighted SNPs. We utilized publicly available annotation tools (VEP and Haploreg) which indicated that four of the highlighted SNPs were in regulatory regions and a substantial fraction (24/56) overlapped enhancer and promoter sites in brain tissue. One SNP, rs13385826 in *ERBB4* which was associated with both SI and ND, was predicted to be a splicing disruption variant. Using in-house blood-derived data (DILGOM), we identified several SNPs that were eQTLs and meQTLs, many of which we could replicate in GTEx and mQTLdb respectively. Since smoking behavior is a neuropsychiatric phenotype, we extended this search to brain tissues availing data from resources like GTEx and BRAINEAC for eQTLs, and meQTLs in fetal brain [212]. We could replicate the majority of our meQTL findings from blood in brain tissue but not all eQTLs, suggesting that smoking associated SNPs drive methylation and expression of NSP genes differently across tissues.

Annotation of the identified variants illustrates their functional potential and highlights the importance of using multi-omics data from the same individual, providing insight into the plausible mechanistic action of non-coding variants in regulating gene function. For example, in the blood-derived DILGOM sample, rs35225704 in *ERBB4* was identified both as an eQTL (major allele decreases expression, Beta=-0.07) and meQTL (major allele increases methylation at cg16329650 (Beta=0.04) but decreases at cg01965462 (Beta=-0.05)). An interplay between genotype, methylation and gene expression at this locus is possible, wherein methylation at these CpG sites may be the mechanism regulating the observed association between expression levels and genotype.

We examined differential expression and methylation of the NSP genes between smokers and never-smokers. A significant difference in expression of *NRG1* and *PSEN1* between the two groups was observed. Interestingly, pooling occasional smokers with current smokers diluted the observed significant difference, suggesting that the NSP expression in occasional smokers resembles that of never-smokers. This also highlights the importance of carefully defining smoking status (current/regular *versus* occasional smokers) in gene expression studies. Unlike gene expression, no significant differences in methylation of the NSP genes were observed, suggesting other mechanisms might be regulating the expression levels.

As NSP is also implicated in SCZ [117], where patients likely self-medicate by smoking [81], we speculated if smoking was confounding the apparent association with the disease. In our small sample of SCZ cases and controls, we observed a trend (P-value<0.1) for differential expression in *NRG1*, *NRG3,* and *APH1B* which became non-significant when smoking status was included as a covariate in the analysis. These results, although not statistically significant (low power), hint at confounding the effect of smoking on the NSP expression in SCZ. This has been neglected in several studies reporting differential expression of the NSP genes in SCZ (data mined from SZDB - Table 6).

In this study, we demonstrated a thorough examination of the NSP applying a hypothesis-driven approach. Utilizing in-house and publicly available data, we uncovered potential functions of the associations identified. Using a phenotypically rich family sample, we provide support to our hypothesis, identifying association between genetic variants in seven of the key NSP genes and three distinct phases of smoking behavior - smoking initiation, nicotine dependence and nicotine withdrawal. We did not detect any association with alcohol-related phenotypes, indicating that this pathway may be specific to smoking and not addiction to other substances, although more studies with other illicit substance use are needed to confirm this. Taken together, our results illustrate the involvement of the NSP in smoking behavior with several potentially functional SNPs which could be probable therapeutic targets for smoking cessation and other psychiatric comorbidities.

6 Challenges and prospects

With the advent of technology, disease gene mapping has seen tremendous progress. Earlier studies examined only a handful of genetic markers; now high-throughput microarrays are common and sequencing technologies are also slowly becoming routine, but we still lack proper annotation and interpretation of the findings. GWAS have led to a better understanding of the polygenic architecture of complex traits, however, there remain several challenges with regards to reaching a clear mechanistic understanding of how complex disease loci influence the phenotype. The majority of GWAS hits explain only a small proportion of the total heritability, with small effect sizes observed for individual loci, leaving a significant portion unexplained. Along with that, the huge burden of multiple hypothesis testing necessitates the need for large samples to be analyzed to provide adequate statistical power to implicate any locus. Further, since most of the identified associations reside in the non-coding region of the genome, the molecular mechanisms by which they might regulate the trait expression remains unexplained.

Biomarkers, such as metabolites, provide a valuable alternative for objective assessment of behavioral phenotypes, given the measurement precision and biological proximity. As they are precise molecular measures of trait variance, they provide higher statistical power to capture association signals even in smaller samples, as opposed to self-reported phenotypes. For example, the difference mentioned by Ware et al. [74] between using cotinine *versus* cigarettes per day in a GWAS provided greater power to identify variants that explain a larger proportion of variance in the phenotype. However, large-scale consortia efforts have enabled meta-analyses of multiple cohorts, gradually overcoming power issues with self-reported or objectively assessed phenotypes identifying more robust associations with precise estimates of effect sizes. This approach has worked well for GWAS for over a decade now, but for EWAS, the variation in the epigenome due to differences in genetic makeup, lifestyle, environmental exposures, age, the tissue being studied, and cellular heterogeneity hinder the analysis and interpretation of results. However, EWAS meta-analyses success stories are emerging [142, 153, 229]. With methylation studies, identifying whether the association is capturing the phenotype of interest or an underlying comorbidity can be challenging. For example, in the recent EWAS meta-analysis of educational attainment levels by Linner et al. [229], the association signals overlapped with most prominent smoking associated loci. On including smoking as a covariate in their analysis, all association signals disappeared (negative correlation between education and smoking [230]); but

no such issue was apparent in the GWAS meta-analysis of educational attainment by the same consortium [231].

Among the studies included in this thesis, sample sizes were modest which could have resulted in false negatives. In Appendix II, power calculations for the studies included in this thesis are included that show we had sufficient power to identify large effects in all three studies but modest power to identify smaller effects. While a larger sample size could potentially identify additional associations, for study I and II, utilizing biomarkers of nicotine metabolism and nicotine exposure allowed us to identify associations in biologically meaningful loci in our samples. The family sample utilized in study III is a valuable source for studying substance use disorders given its rich phenotype profile capturing a broad range of substance use behaviors. Family samples have certain advantages as well as limitations in statistical analysis. While a smaller sample of family design is sufficient in identifying rare inherited variants, but at the same time, large pedigrees are essential for studying the passage of disease loci. In the NAG-FIN sample, although there were 740 families, only a handful of families had large pedigrees. Since at the time of the sample collection the mean age of the twins was 57 years, few parents could participate. Incomplete family structures or smaller pedigrees may decrease power in family-based analysis. However, we used PSEUDOMARKER for joint linkage and LD analysis, which is evidenced to be the most powerful tool for family-based analysis [232] and can handle missing parents and genotypes. It extracts maximum information from a family sample enabling detection of significant associations in seven of the ten NSP genes. The NAG-FIN sample, although ascertained for heavy smoking individuals, did not have biochemical verification of the smoking status (unlike in study I and II). However, it is unlikely that many non-smokers would have claimed to be smokers in the self-report questionnaire; they consistently reported being a smoker in multiple follow-ups data collection.

Several avenues are important for further research to improve interpretation of findings from association studies. Using a higher resolution of the genome, integration of multi-omics datasets to assess mediation and causality and profiling the omics data in multiple tissues would be highly beneficial understanding the variation in traits more extensively. A potential limitation of our studies was the use of blood-derived methylation profiles when it is well known that tissue-specific DNA-methylation profiles exist. Methylation profiles in blood might not be entirely representative of tissue of interest, which in our case would have been liver where nicotine is primarily metabolized [68] for study I and II, and brain for study III. However, blood is more readily

accessible (for clinical purposes as well) and functional than other relevant target tissues. For example, blood cells are the first to absorb the effects of smoke exposure in lungs and circulate throughout the body. GTEx is an excellent example of cross-tissue expression profiles. More such resources, including multiple layers of omics data from the same individual across different tissues, could be valuable to the scientific community. It would enable translation of mere associations into a clearer picture of functional mechanisms and downstream consequences. Adding extra layers of omics data may not be as easy as it sounds. Except for genetic data, other omics data may not be stable across time or cell cycle. For instance, expression data in blood is not identical at two different time-points at all loci, even if the blood samples have been drawn from an individual on the same day.

Given that most complex diseases are influenced by both genetic and environmental factors, to have a full mechanistic insight of how they develop over time would require coordinated sets of several omics data at multiple time points, collected from many disease-relevant tissues. We examined only three of such data layers - genetic, epigenetic and expression profiles. There could potentially be many more intermediary mechanisms affecting gene regulation which we did not capture. It should also be noted that in this thesis, we have only looked at one of the several epigenetic modifiers, DNA methylation. Other epigenetic mechanisms may also contribute toward the observed variance in our phenotypes. For example, microRNA 101 targets two key genes identified in study II (EWAS on cotinine) - *AHRR* and *CYP2C18* [233]. The exact mechanism by which this microRNA alters gene activity remains to be elucidated. Similarly, exploring the role of other structural genetic variations including ultra-rare SNPs and copy number variations might also contribute toward the missing heritability of our phenotypes.

Causality is difficult to prove but would be greatly informative in understanding disease development process and treatment design. Methylation changes that are a consequence of an exposure prove to be useful biomarkers, for example, smoking-related methylation changes can predict disease risk, like in case of lung cancer [234], Multiple Sclerosis [235] and all-cause mortality [236]. Distinguishing methylation changes that lie in the causal pathway from those that are a consequence of the disease is crucial for understanding disease etiology. Combining genetic and epigenetic data in a causal inference test scenario can be helpful in identifying if methylation lies on the causal pathway between the association of a trait and genetic variation. We utilized the CIT to assess causal mediation in our studies. CIT works well for assessing mediation, as it is based on an initial hypothesis of a causal model (genotype $\rightarrow$ mediator $\rightarrow$

phenotype; assuming no other causal link between genotype and phenotype) and hence is free from problems of reverse causation and confounding [237, 238]. However, the limitation of CIT comes from its inability to handle more complex causal models with more than one molecular mediator, which might be useful if more biological data layers are available. More flexible methods which explore the causal structure from the data after evaluating most plausible structures, such as structural equation modeling or Bayesian framework based tools may be useful for such scenarios.

Microarray technology has afforded huge progress in the availability of omics data because of its low-cost high-throughput output. However, coverage of the genome with microarray varies significantly between different omics layers at present. For instance, genotyping arrays with imputation cover ~90% of the genome [239], expression arrays only cover a few exons per gene, and the methylation array used in this thesis (450k) covers ~1% of the CpG sites in the human genome. Such inconsistent coverage resulting in an incomplete and biased view of the relative contribution of genetic and epigenetic factors to trait variation may be a limitation such that not all disease-relevant loci can be identified. Sequencing technology, for example, whole genome sequencing for genetic data, RNA-seq for expression data and whole-genome bisulfite sequencing for methylation data would provide a much higher resolution into the genome. Although implementing all these newer technologies to dig deeper into the genome hold promise, they come with additional costs and the added burden of multiple hypothesis testing. It is a matter of what the research question is, screening *versus* targeting a region in a magnified fashion. Other enhancement avenues using microarrays include the use of population-specific imputation reference panels [240] for genotype data and higher coverage EPIC array (the successor of the 450k) for methylation data.

Studies included in this thesis were performed in Finnish population samples. Finnish population, a genetic isolate, has already proven extremely useful for mapping genes involved in rare monogenic disorders [241] but also suggests that the results may be specific to Finnish population. This supports the idea of designing personalized treatments. For instance, genetic data can be used to determine the efficacy of a drug/medication or its dosage, possibly preventing adverse drug responses and in-turn saving costs for treatment facilities. Such genetic profiling was tested in a clinical trial to examine efficacy of smoking cessation treatments [105]. Participant's nicotine metabolism rate was characterized using a genotype-based metric (*CYP2A6* alleles [242, 243]) and the results indicate that slow metabolizers do not benefit from active

treatment (nicotine replacement therapy, bupropion, or combination) compared to placebo, while normal or fast metabolizers benefit significantly [105]. This means individuals with genetically determined slow nicotine metabolism rate, can avoid taking ineffective treatments and potentially their toxic effects. Such patient-specific prescriptions may also benefit from the use of other omics data, thereby leading to a more accurate medication prescription and dosage tailored to individual patients. Candidate pathways and genes and/or their downstream targets also present potential targets for smoking cessation aid [244] and smoking comorbid psychiatric illnesses. For instance, downstream targets of the NSP are being evaluated as drug targets in SCZ [245, 246]. Epigenetic alterations represent new vistas for therapeutic development and selective manipulation of epigenetic marks such as methylation in genes like *CYP2A6* and *CYP2C18* involved in metabolism of xenobiotics hold high potential [247]. It is only once causal genes are identified that designing personalized therapeutics comes into picture. Actual translation of the identified loci into tangible targets remains a major challenge in solving the complex puzzle of smoking.

7 Conclusion

One of the key avenues in public health is the identification of modifiable factors that causally influence disease risk. Owing to the adverse health effects of nicotine and tobacco, understanding the consequences and mechanisms involved in nicotine exposure are crucial. With technological advancements, we can dig deeper into the genome and identify susceptibility loci implicated in complex traits. Identifying and characterizing the contribution of disease-risk associated variants is crucial in taking healthcare toward personalized medicine. The findings presented in this thesis are centered on the theme of genetic and epigenetic alterations affecting smoking behavior via genes involved in nicotine metabolism, nicotine dependence, and neuronal pathways.

Study I and II showed the utility of a hypothesis-free scan of the genome using biomarkers of smoking, revealing genetic loci involved in nicotine metabolism and epigenetically regulated smoking-related genes. In both studies, an interplay between the genome and epigenome, and evidence of methylation causally mediating some of these effects was noted. Probable downstream effects of identified associations were examined through the integration of multi-omics datasets. These two studies also highlight the benefit of using a biomarker as a phenotype, providing greater statistical power as well as biological proximity to the exposure, in identifying relevant loci for behavioral traits. In study III, using a targeted approach we validated the role of neuregulin signaling pathway specifically in smoking behavior and highlighted functional potential in most of variants identified through integrative analysis of multi-omics datasets. Our results also suggest confounding effects of smoking should be taken into consideration for gene expression studies, especially for highly comorbid disorders like schizophrenia.

In conclusion, the work presented in this thesis provides insight into the association and interplay of genetic and epigenetic variants that influence the complex trait of smoking behavior. Further, using multiple biological data layers the regulatory potential of seemingly non-functional trait associated variation was assessed. Our findings suggest that the genome and the epigenome act in concert and methylation may be a molecular mechanism mediating the observed effects in some genes. The research produced in this thesis is not only relevant for public health but also highly crucial for meeting the bigger goal of making the world nicotine and tobacco free. It opens a window of opportunity in genomics-based personalization of pharmacotherapeutics for enhanced smoking cessation and treatment of comorbid disorders.

Acknowledgments

This Ph.D. work was carried out at the Department of Public Health (Hjelt Institute) and Institute for Molecular Medicine, University of Helsinki and financially supported by the Marie Curie initial training network- EpiTrain, Academy of Finland, University of Helsinki Research Funds, and Sigrid Juselius Foundation grants to my supervisors Prof. Jaakko Kaprio and Adj. Prof. Miina Ollikainen.

I would like to thank Jaakko and Miina for giving me the opportunity to pursue this academic journey, in a country far from home. Came with a lot of surprises. Jaakko, thanks for letting me be a part of your warm and welcoming group. Thanks for being so approachable and always sharing your wisdom. I have learned a lot and grown both, personally and professionally. Miina, thanks for providing me with the guidance and encouragement to be independent. Thanks for the support and understanding you've shown throughout. Adj. Prof. Anu Loukola, thank you for playing such a huge and crucial part in this journey. I have learned so much from you, not just about research but much more. Your enthusiasm for science is contagious.

Adj. Prof. Bas Heijmans, I would like to thank you for accepting the position as the official opponent at the public examination of this thesis. I would also like to thank the pre-examiners, Dr. Eilis Hannon and Adj. Prof. Harri Lähdesmäki, for providing constructive feedback on the thesis making it much more concrete. My thesis committee members - Adj. Prof. Sippy Kaur, thanks for providing me a home away from home - much love to the little ones and Dr. Simon Anders, for shedding light on the prospects of my work and asking tough, yet important questions.

A special note of gratitude to my awesome, supportive and fun colleagues Aileen, Aline, Alyce, Anu R, Beenish, Emma, Jade, Jenni, Lalitha, Mahes, Pauliina, Teemu and many others not mentioned here for your friendship and support, for making the little Finnish adventures come true, for countless lunches and the warm heartfelt surprises and hugs. I'll always cherish my friendship with each one of you.

Dearest Papa, Ma, and Annu, saying thanks is just not enough to express how grateful I am for all your love and support. I wouldn't be here if it weren't for you. Thanks, Sakshi Bhabhi for all the uplifting little chats and taking care of the loved ones. Last but certainly not the least, Anuj thanks for all the patience you've had and the support you've given me, without you this would've been a lot harder. I cannot thank you all enough.

Appendix I

Results from study II. Association between serum cotinine levels of regular smokers and DNA methylation. Sensitivity analysis performed to ascertain the effect of using a lower threshold (Cotinine>4.85ng/ml; N=310) to define regular smokers in contrast to the standard threshold of cotinine>10 ng/ml (N=293).

CpG	Chr	Nearest gene	Discovery Analysis				Secondary Analysis [a]				Sensitivity analysis [b]			References
			Beta	SE	P value	P_{FDR}	Beta	SE	*P*-value	P_{FDR}	Beta	SE	*P*-value	
cg25189904	1	*GNG12*	-0.000135	0.000026	5.6E-07	9.0E-03	-0.000134	0.000028	1.8E-06	3.1E-02	-0.000126	0.000031	5.6E-05	[125, 128, 130-133, 135, 139-144]
cg21033965	1	*CLEC20A*	-0.000096	0.000020	4.0E-06	3.7E-02	**				-0.000092	0.000023	8.3E-05	[142]
cg26687670	1	*OPN3*	-0.000025	0.000005	6.5E-06	4.4E-02	**				-0.000022	0.000006	3.4E-04	Novel
cg27241845	2	*ALPP*	-0.000083	0.000016	3.3E-07	6.2E-03	-0.000082	0.000016	8.8E-07	1.7E-02	-0.000076	0.000018	3.9E-05	[128, 130, 132, 133, 135, 139, 141, 142, 144]
cg03329539	2	*ALPPL2*	-0.000070	0.000012	2.7E-08	6.6E-04	-0.000069	0.000013	1.3E-07	3.0E-03	-0.000080	0.000014	2.4E-08	[128, 130-133, 135, 139-142, 144]
cg06644428	2	*ALPPL2*	-0.000054	0.000012	5.5E-06	4.3E-02	**				-0.000053	0.000013	7.7E-05	[125, 128, 130, 132-135, 139-142, 144]
cg05951221	2	*ALPPL2*	-0.000178	0.000021	1.9E-15	2.0E-10	-0.000171	0.000022	1.4E-13	1.5E-08	-0.000177	0.000024	2.7E-12	[125, 128, 130-135, 137, 139-144]
cg21566642	2	*ALPPL2*	-0.000210	0.000026	1.1E-14	8.8E-10	-0.000205	0.000027	4.2E-13	3.4E-08	-0.000195	0.000030	2.1E-10	[125, 128, 130-135, 137, 139-142, 144]
cg01940273	2	*ALPPL2*	-0.000134	0.000017	9.8E-14	6.3E-09	-0.000129	0.000018	3.4E-12	2.2E-07	-0.000119	0.000020	5.9E-09	[125, 128, 130-133, 135, 137, 139-144]
cg13193840	2	*ALPPL2*	-0.000045	0.000010	3.6E-06	3.4E-02	-0.000048	0.000010	2.5E-06	3.3E-02	-0.000044	0.000011	7.4E-05	[128, 132, 133, 135, 139, 141, 142, 144]
cg02306995	3	*SEMA5B*	*				0.000047	0.000010	4.4E-06	4.8E-02	0.000040	0.000011	4.3E-04	Novel ***
cg04776445	4	*MSX1*	*				0.000085	0.000018	4.3E-06	4.8E-02	0.000076	0.000020	1.6E-04	Novel
cg13740236	4	*LSM6*	-0.000053	0.000009	2.9E-08	6.7E-04	-0.000057	0.000010	1.0E-08	3.0E-04	-0.000047	0.000011	1.2E-05	Novel
cg11902777	5	*AHRR*	-0.000071	0.000011	1.1E-09	3.2E-05	-0.000068	0.000012	2.1E-08	5.6E-04	-0.000077	0.000013	9.9E-09	[128, 132, 135, 139, 141, 144]
cg01899089	5	*AHRR*	-0.000063	0.000014	6.1E-06	4.4E-02	**				-0.000072	0.000016	5.1E-06	[128, 130, 132, 135, 141, 142, 144]
cg05575921	5	*AHRR*	-0.000359	0.000039	3.3E-18	1.1E-12	-0.000351	0.000041	4.4E-16	1.1E-10	-0.000353	0.000045	6.3E-14	[125, 126, 128-137, 139-144]
cg26703534	5	*AHRR*	-0.000086	0.000013	5.8E-10	1.9E-05	-0.000084	0.000014	6.2E-09	2.0E-04	-0.000076	0.000015	1.4E-06	[128-132, 135, 136, 140-142, 144]
cg14817490	5	*AHRR*	-0.000099	0.000021	2.4E-06	2.9E-02	**				-0.000098	0.000024	5.3E-05	[128, 130-133, 135, 136, 139-144]
cg25648203	5	*AHRR*	-0.000091	0.000015	4.3E-09	1.1E-04	-0.000086	0.000016	1.1E-07	2.8E-03	-0.000085	0.000017	1.4E-06	[125, 128, 130-133, 135, 139-144]
cg21161138	5	*AHRR*	-0.000142	0.000016	1.8E-17	2.9E-12	-0.000141	0.000017	6.6E-16	1.1E-10	-0.000138	0.000018	4.0E-13	[125, 128-133, 135, 136, 139-142, 144]
cg24090911	5	*AHRR*	-0.000076	0.000015	4.3E-07	7.4E-03	-0.000074	0.000015	2.5E-06	3.3E-02	-0.000067	0.000016	4.8E-05	[128, 132, 133, 135, 141, 142, 144]
cg10961758	5	*EDIL3*	*				-0.000069	0.000014	2.5E-06	3.3E-02	-0.000040	0.000016	1.1E-02	Novel***
cg16179182	5	*VTRNA1-1*	0.000048	0.000010	5.7E-06	4.3E-02	**				0.000051	0.000012	3.6E-05	Novel
cg14580211	5	*SMIM3*	-0.000062	0.000013	1.7E-06	2.2E-02	**				-0.000046	0.000015	1.6E-03	[128, 130, 132, 135, 139-142, 144]
cg06126421	6	*IER3*	-0.000132	0.000020	1.1E-10	4.3E-06	-0.000132	0.000021	6.9E-10	2.8E-05	-0.000123	0.000023	9.5E-08	[125, 128, 130-133, 135, 137, 139-142, 144]
cg14753356	6	*IER3*	-0.000072	0.000015	3.1E-06	3.4E-02	**				-0.000072	0.000017	3.8E-05	[128, 130, 132, 135, 139, 141, 142, 144]
cg22856972	6	*TBPL1*	0.000066	0.000014	3.0E-06	3.4E-02	0.000073	0.000014	9.1E-07	1.7E-02	0.000061	0.000016	1.4E-04	Novel
cg05469934	7	*PDGFA*	0.000038	0.000008	3.6E-06	3.4E-02	**				0.000038	0.000009	1.1E-05	Novel
cg09022230	7	*TNRC18*	-0.000079	0.000014	4.1E-08	8.8E-04	-0.000070	0.000014	2.4E-06	3.3E-02	-0.000078	0.000016	2.4E-06	[130, 132, 135, 141, 142, 144]
cg21322436	7	*CNTNAP2*	-0.000055	0.000012	5.6E-06	4.3E-02	**				-0.000045	0.000014	9.8E-04	[128, 132, 135, 139, 141, 142, 144]
cg25949550	7	*CNTNAP2*	-0.000045	0.000009	1.4E-06	1.9E-02	**				-0.000049	0.000011	5.0E-06	[125, 128, 130, 132, 135, 139-142, 144]
cg09267815	8	*YTHDF3*	-0.000080	0.000016	7.2E-07	1.1E-02	**				-0.000084	0.000018	5.4E-06	Novel

cg22288066	8	*NOV*	0.000033	0.000007	6.8E-06	4.4E-02	**				0.000037	0.000008	8.9E-06	Novel
cg24838345	8	*MTSS1*	-0.000088	0.000017	2.6E-07	5.3E-03	-0.000084	0.000017	2.3E-06	3.3E-02	-0.000078	0.000019	6.5E-05	[130, 132, 135, 139, 141, 142, 144]
cg25305703	8	*CASC8*	-0.000088	0.000018	1.7E-06	2.2E-02	**				-0.000090	0.000021	2.3E-05	[128, 132, 135, 141, 142, 144]
cg01692968	9	*SLC44A1*	-0.000075	0.000016	3.3E-06	3.4E-02	-0.000079	0.000017	2.7E-06	3.3E-02	-0.000064	0.000018	3.7E-04	[128, 130, 132, 135, 139, 141, 142, 144]
cg14120703	9	*NOTCH1*	-0.000043	0.000009	4.3E-06	3.8E-02	-0.000045	0.000010	4.3E-06	4.8E-02	-0.000042	0.000010	7.7E-05	[135, 142]
cg05992400	10	*CYP2C18*	-0.000050	0.000011	4.9E-06	4.1E-02	**				-0.000047	0.000012	1.6E-04	Novel
cg16421726	10	*ZRANB1*	0.000056	0.000012	2.7E-06	3.1E-02	**				0.000057	0.000013	1.8E-05	Novel
cg05335388	11	*ABTB2*	*				0.000035	0.000007	1.6E-06	2.9E-02	0.000030	0.000008	2.2E-04	Novel***
cg19254163	11	*PTGDR2*	-0.000042	0.000009	6.4E-06	4.4E-02	**				-0.000043	0.000010	4.0E-05	[128, 132, 135, 141, 142, 144]
cg21611682	11	*LRP5*	-0.000071	0.000010	1.2E-12	6.7E-08	-0.000067	0.000010	8.2E-11	3.8E-06	-0.000072	0.000011	3.5E-10	[125, 128, 130, 132, 133, 135, 139-142, 144]
cg14624207	11	*LRP5*	-0.000063	0.000010	2.6E-10	9.2E-06	-0.000063	0.000010	1.4E-09	5.1E-05	-0.000060	0.000011	4.4E-08	[128, 132, 135, 139, 141, 142, 144]
cg10788371	11	*LRRC32*	-0.000064	0.000013	1.3E-06	1.8E-02	-0.000067	0.000013	8.6E-07	1.7E-02	-0.000069	0.000015	5.5E-06	[132, 135, 141, 142, 144]
cg11491407	12	*CACNA2D4*	-0.000095	0.000020	5.7E-06	4.3E-02	**				-0.000098	0.000024	4.4E-05	Novel
cg18443392	12	*LRP1*	0.000027	0.000006	4.2E-06	3.8E-02	**				0.000025	0.000006	1.4E-04	Novel
cg04524475	13	*RNY1P5*	-0.000053	0.000012	6.2E-06	4.4E-02	**				-0.000041	0.000013	2.3E-03	Novel
cg26589665	15	*THSD4*	-0.000042	0.000008	3.8E-07	6.9E-03	-0.000041	0.000008	2.1E-06	3.3E-02	-0.000040	0.000009	1.6E-05	Novel
cg05361530	16	*GNG13*	0.000044	0.000009	4.8E-06	4.1E-02	**				0.000031	0.000010	3.3E-03	Novel
cg07991586	16	*ADCY7*	-0.000042	0.000009	6.6E-06	4.4E-02	**				-0.000034	0.000011	1.2E-03	Novel
cg05460226	17	*PIK3R5*	-0.000072	0.000015	3.4E-06	3.4E-02	**				-0.000082	0.000017	3.0E-06	[135, 141, 142]
cg05396243	17	*TMEM220-AS1*	-0.000036	0.000007	1.0E-06	1.5E-02	**				-0.000028	0.000008	7.0E-04	Novel
cg07658936	17	*ARHGAP44*	0.000029	0.000006	6.8E-06	4.4E-02	**				0.000021	0.000007	2.7E-03	Novel
cg03636183	19	*F2RL3*	-0.000151	0.000022	1.9E-11	8.7E-07	-0.000155	0.000022	2.8E-11	1.5E-06	-0.000145	0.000025	1.6E-08	[122, 123, 125, 127, 128, 130-133, 135-144]
cg12817959	22	*SSTR3*	*				0.000034	0.000007	4.3E-06	4.8E-02	0.000028	0.000008	3.7E-04	Novel***

** Probe not significant in discovery analysis;*

*** Probe not significant in the secondary analysis;*

****Novel hits identified in the secondary analysis.*

[a] *EWAS including the GRS for NMR based on independent SNPs identified in study I.*

[a] *Sensitivity analysis for top 55 CpG sites in a sample of regular smoker based on a threshold of cotinine>10ng/ml (N=293)*

Appendix II

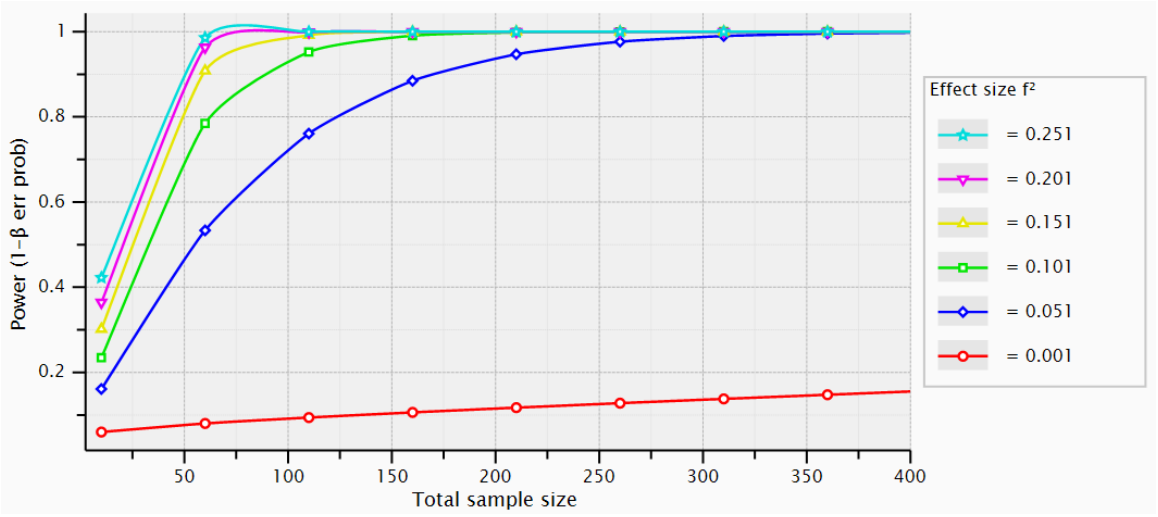
Power Calculations

Study I: We estimated the power of our GWAS meta-analysis sample to detect signals at the P-value<5E-8 significance threshold assuming an LD (r2) of 0.8 between the causal allele and the marker allele. We used the Genetic Power Calculator [248] to estimate the power.

	MAF 0.01	**MAF 0.03**	**MAF 0.05**	**MAF 0.1**	**MAF 0.2**	**MAF 0.4**
Beta ±0.1	0.00%	0.00%	0.00%	0.00%	0.02%	0.11%
Beta ±0.2	0.00%	0.01%	0.04%	0.58%	6.17%	25.61%
Beta ±0.3	0.00%	0.16%	1.20%	14.72%	66.79%	**96.23%**
Beta ±0.4	0.02%	1.71%	11.75%	66.79%	**99.25%**	**100.00%**
Beta ±0.5	0.13%	10.02%	45.53%	**97.32%**	**100.00%**	**100.00%**
Beta ±0.6	0.56%	33.02%	**83.38%**	**99.97%**	**100.00%**	**100.00%**
Beta ±0.7	2.03%	65.72%	**98.02%**	**100.00%**	**100.00%**	**100.00%**
Beta ±0.8	5.94%	**89.50%**	**99.92%**	**100.00%**	**100.00%**	**100.00%**

According to the power analyses we had inadequate power to detect signals with rare SNPs (MAF <5%) unless they have very large effect sizes, but high power to detect signals with common SNPs (MAF >5%) that have medium to high effect sizes (Beta>±0.6).

Study II: Using the G*Power tool [249], we performed *post-hoc* power analysis with a significance level of 0.05 and sample size of 310, we have high power to detect large effects (0.3-0.05), however, we have low power to detect smaller effect sizes (~0.001).



Study III: Similar to study I, we estimated the power of our association discovery sample (NAG-FIN; N=1998) to detect signals at the 3.5E-04 significance threshold (equivalent to FDR adjusted P-value=0.05) assuming an LD (r2) of 0.8 between the causal allele and the marker allele using the Genetic Power Calculator [248]:

	MAF 0.01	**MAF 0.03**	**MAF 0.05**	**MAF 0.1**	**MAF 0.2**	**MAF 0.4**
Beta ±0.1	0.13 %	0.45 %	0.96 %	3.02 %	9.47 %	21.11 %
Beta ±0.2	0.71 %	5 %	13.38 %	42.94 %	**83.14** %	**97.65** %
Beta ±0.3	2.96 %	24.87 %	55.09 %	**93.7** %	**99.94** %	**100** %
Beta ±0.4	9.28 %	61.41 %	**91.45** %	**99.94** %	**100** %	**100** %
Beta ±0.5	22.38 %	**89.64** %	**99.55** %	**100** %	**100** %	**100** %
Beta ±0.6	42.27 %	**98.73** %	**99.99** %	**100** %	**100** %	**100** %
Beta ±0.7	64.45 %	**99.93** %	**100** %	**100** %	**100** %	**100** %
Beta ±0.8	**82.56** %	**100** %	**100** %	**100** %	**100** %	**100** %

Our discovery sample had adequate power (>80%) to detect variants with MAF>0.05 and medium to large effect sizes (Beta>±0.5), and low power to detect variants with MAF<0.01 with similar effect sizes.

References

1. Tenesa, A. and C.S. Haley, *The heritability of human disease: estimation, uses and abuses.* Nat Rev Genet, 2013. **14**(2): p. 139-49.
2. Polderman, T.J., et al., *Meta-analysis of the heritability of human traits based on fifty years of twin studies.* Nat Genet, 2015. **47**(7): p. 702-9.
3. Visscher, P.M., W.G. Hill, and N.R. Wray, *Heritability in the genomics era--concepts and misconceptions.* Nat Rev Genet, 2008. **9**(4): p. 255-66.
4. Frazer, K.A., et al., *Human genetic variation and its contribution to complex traits.* Nat Rev Genet, 2009. **10**(4): p. 241-51.
5. MacArthur, J., et al., *The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog).* Nucleic Acids Res, 2017. **45**(D1): p. D896-D901.
6. Patnala, R., J. Clements, and J. Batra, *Candidate gene association studies: a comprehensive guide to useful in silicotools.* BMC Genetics, 2013. **14**(1): p. 39.
7. Tabor, H.K., N.J. Risch, and R.M. Myers, *Candidate-gene approaches for studying complex genetic traits: practical considerations.* Nat Rev Genet, 2002. **3**(5): p. 391-7.
8. Caporaso, N., et al., *Genome-wide and candidate gene association study of cigarette smoking behaviors.* PLoS One, 2009. **4**(2): p. e4653.
9. Fowler, J.S., et al., *Monoamine oxidase and cigarette smoking.* Neurotoxicology, 2003. **24**(1): p. 75-82.
10. Gusev, A., et al., *Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases.* Am J Hum Genet, 2014. **95**(5): p. 535-52.
11. Caballero, A., A. Tenesa, and P.D. Keightley, *The Nature of Genetic Variation for Complex Traits Revealed by GWAS and Regional Heritability Mapping Analyses.* Genetics, 2015. **201**(4): p. 1601-13.
12. Belsky, D.W., et al., *Polygenic risk and the developmental progression to heavy, persistent smoking and nicotine dependence: evidence from a 4-decade longitudinal study.* JAMA Psychiatry, 2013. **70**(5): p. 534-42.
13. Meyers, J.L., et al., *Interaction between polygenic risk for cigarette use and environmental exposures in the Detroit Neighborhood Health Study.* Transl Psychiatry, 2013. **3**: p. e290.
14. Bomba, L., K. Walter, and N. Soranzo, *The impact of rare and low-frequency genetic variants in common disease.* Genome Biology, 2017. **18**.
15. Manolio, T.A., et al., *Finding the missing heritability of complex diseases.* Nature, 2009. **461**(7265): p. 747-53.
16. Wang, S.R., et al., *Simulation of Finnish population history, guided by empirical genetic data, to assess power of rare-variant tests in Finland.* Am J Hum Genet, 2014. **94**(5): p. 710-20.
17. Trerotola, M., et al., *Epigenetic inheritance and the missing heritability.* Hum Genomics, 2015. **9**: p. 17.
18. Handy, D.E., R. Castro, and J. Loscalzo, *Epigenetic modifications: basic mechanisms and role in cardiovascular disease.* Circulation, 2011. **123**(19): p. 2145-56.
19. Wagner, J.R., et al., *The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts.* Genome Biol, 2014. **15**(2): p. R37.
20. Weinhold, B., *Epigenetics: the science of change.* Environ Health Perspect, 2006. **114**(3): p. A160-7.
21. Paul, D.S., N. Soranzo, and S. Beck, *Functional interpretation of non-coding sequence variation: concepts and challenges.* Bioessays, 2014. **36**(2): p. 191-9.

22. Stunnenberg, H.G., C. International Human Epigenome, and M. Hirst, *The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery.* Cell, 2016. **167**(7): p. 1897.
23. Roadmap Epigenomics, C., et al., *Integrative analysis of 111 reference human epigenomes.* Nature, 2015. **518**(7539): p. 317-30.
24. Consortium, E.P., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.
25. Cazaly, E., et al., *Genetic Determinants of Epigenetic Patterns: Providing Insight into Disease.* Mol Med, 2015. **21**: p. 400-9.
26. Rakyan, V.K., et al., *Epigenome-wide association studies for common human diseases.* Nat Rev Genet, 2011. **12**(8): p. 529-41.
27. Alexander, R.P., et al., *Annotating non-coding regions of the genome.* Nat Rev Genet, 2010. **11**(8): p. 559-71.
28. Zhang, F. and J.R. Lupski, *Non-coding genetic variants in human disease.* Hum Mol Genet, 2015. **24**(R1): p. R102-10.
29. Tak, Y.G. and P.J. Farnham, *Making sense of GWAS: using epigenomics and genome engineering to understand the functional relevance of SNPs in non-coding regions of the human genome.* Epigenetics Chromatin, 2015. **8**: p. 57.
30. Edwards, S.L., et al., *Beyond GWASs: illuminating the dark road from association to function.* Am J Hum Genet, 2013. **93**(5): p. 779-97.
31. Huang, Q., *Genetic study of complex diseases in the post-GWAS era.* J Genet Genomics, 2015. **42**(3): p. 87-98.
32. Gutierrez-Arcelus, M., et al., *Passive and active DNA methylation and the interplay with genetic variation in gene regulation.* Elife, 2013. **2**: p. e00523.
33. Albert, F.W. and L. Kruglyak, *The role of regulatory variation in complex traits and disease.* Nat Rev Genet, 2015. **16**(4): p. 197-212.
34. Millstein, J., et al., *Disentangling molecular relationships with a causal inference test.* BMC Genet, 2009. **10**: p. 23.
35. Sheehan, N.A., et al., *Mendelian randomisation and causal inference in observational epidemiology.* PLoS Med, 2008. **5**(8): p. e177.
36. Genomes Project, C., et al., *A global reference for human genetic variation.* Nature, 2015. **526**(7571): p. 68-74.
37. International HapMap, C., *The International HapMap Project.* Nature, 2003. **426**(6968): p. 789-96.
38. McCarthy, S., et al., *A reference panel of 64,976 haplotypes for genotype imputation.* Nat Genet, 2016. **48**(10): p. 1279-83.
39. Bock, C., *Analysing and interpreting DNA methylation data.* Nat Rev Genet, 2012. **13**(10): p. 705-19.
40. Touleimat, N. and J. Tost, *Complete pipeline for Infinium((R)) Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation.* Epigenomics, 2012. **4**(3): p. 325-41.
41. Marabita, F., et al., *An evaluation of analysis pipelines for DNA methylation profiling using the Illumina HumanMethylation450 BeadChip platform.* Epigenetics, 2013. **8**(3): p. 333-46.
42. Morris, T.J. and S. Beck, *Analysis pipelines and packages for Infinium HumanMethylation450 BeadChip (450k) data.* Methods, 2015. **72**: p. 3-8.
43. Cazaly, E., et al., *Comparison of pre-processing methodologies for Illumina 450k methylation array data in familial analyses.* Clin Epigenetics, 2016. **8**: p. 75.
44. Sun, Z., et al., *Base resolution methylome profiling: considerations in platform selection, data preprocessing and analysis.* Epigenomics, 2015. **7**(5): p. 813-28.

45. Wilhelm-Benartzi, C.S., et al., *Review of processing and analysis methods for DNA methylation array data.* Br J Cancer, 2013. **109**(6): p. 1394-402.
46. Wright, M.L., et al., *Establishing an analytic pipeline for genome-wide DNA methylation.* Clin Epigenetics, 2016. **8**: p. 45.
47. Pidsley, R., et al., *A data-driven approach to preprocessing Illumina 450K methylation array data.* BMC Genomics, 2013. **14**: p. 293.
48. Wu, M.C., et al., *A systematic assessment of normalization approaches for the Infinium 450K methylation platform.* Epigenetics, 2014. **9**(2): p. 318-29.
49. Laird, P.W., *Principles and challenges of genomewide DNA methylation analysis.* Nat Rev Genet, 2010. **11**(3): p. 191-203.
50. Hu, J. and X. He, *Enhanced quantile normalization of microarray data to reduce loss of information in gene expression profiles.* Biometrics, 2007. **63**(1): p. 50-9.
51. Teschendorff, A.E., et al., *A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data.* Bioinformatics, 2013. **29**(2): p. 189-96.
52. Maksimovic, J., L. Gordon, and A. Oshlack, *SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips.* Genome Biol, 2012. **13**(6): p. R44.
53. Aryee, M.J., et al., *Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays.* Bioinformatics, 2014. **30**(10): p. 1363-9.
54. Lehne, B., et al., *A coherent approach for analysis of the Illumina HumanMethylation450 BeadChip improves data quality and performance in epigenome-wide association studies.* Genome Biol, 2015. **16**: p. 37.
55. Fortin, J.P., et al., *Functional normalization of 450k methylation array data improves replication in large cancer studies.* Genome Biol, 2014. **15**(12): p. 503.
56. WHO, *WHO global report on trends in prevalence of tobacco smoking*. 2015: Geneva.
57. CDC, *How Tobacco Smoke Causes Disease: The Biology and Behavioral Basis for Smoking-Attributable Disease: A Report of the Surgeon General*. 2010, Atlanta (GA).
58. Aloise-Young, P.A., J.W. Graham, and W.B. Hansen, *Peer influence on smoking initiation during early adolescence: a comparison of group members and group outsiders.* J Appl Psychol, 1994. **79**(2): p. 281-7.
59. Conrad, K.M., B.R. Flay, and D. Hill, *Why children start smoking cigarettes: predictors of onset.* Br J Addict, 1992. **87**(12): p. 1711-24.
60. Mercken, L., et al., *Social influence and selection effects in the context of smoking behavior: changes during early and mid adolescence.* Health Psychol, 2009. **28**(1): p. 73-82.
61. Buller, D.B., et al., *Understanding factors that influence smoking uptake.* Tob Control, 2003. **12 Suppl 4**: p. IV16-25.
62. Allen, S.S., et al., *Craving, withdrawal, and smoking urges on days immediately prior to smoking relapse.* Nicotine Tob Res, 2008. **10**(1): p. 35-45.
63. Ashare, R.L., M. Falcone, and C. Lerman, *Cognitive function during nicotine withdrawal: Implications for nicotine dependence treatment.* Neuropharmacology, 2014. **76 Pt B**: p. 581-91.
64. Association, A.P. and A.P. Association, *Diagnostic and statistical manual of mental disorders (DSM-IV).* Washington, DC: Author, 1994.
65. Heatherton, T.F., et al., *The Fagerstrom Test for Nicotine Dependence: a revision of the Fagerstrom Tolerance Questionnaire.* Br J Addict, 1991. **86**(9): p. 1119-27.
66. Benowitz, N.L., *Cotinine as a biomarker of environmental tobacco smoke exposure.* Epidemiol Rev, 1996. **18**(2): p. 188-204.
67. Baker, T.B., et al., *DSM criteria for tobacco use disorder and tobacco withdrawal: a critique and proposed revisions for DSM-5.* Addiction, 2012. **107**(2): p. 263-75.

68. Sobkowiak, R. and A. Lesicki, *[Absorption, metabolism and excretion of nicotine in humans].* Postepy Biochem, 2013. **59**(1): p. 33-44.
69. Benowitz, N.L., J. Hukkanen, and P. Jacob, 3rd, *Nicotine chemistry, metabolism, kinetics and biomarkers.* Handb Exp Pharmacol, 2009(192): p. 29-60.
70. Strasser, A.A., et al., *Nicotine metabolite ratio predicts smoking topography and carcinogen biomarker level.* Cancer Epidemiol Biomarkers Prev, 2011. **20**(2): p. 234-8.
71. Ray, R., R.F. Tyndale, and C. Lerman, *Nicotine dependence pharmacogenetics: role of genetic variation in nicotine-metabolizing enzymes.* J Neurogenet, 2009. **23**(3): p. 252-61.
72. Kettunen, J., et al., *Genome-wide association study identifies multiple loci influencing human serum metabolite levels.* Nat Genet, 2012. **44**(3): p. 269-76.
73. Demirkan, A., et al., *Insight in genome-wide association of metabolite quantitative traits by exome sequence analyses.* PLoS Genet, 2015. **11**(1): p. e1004835.
74. Ware, J.J., et al., *Genome-Wide Meta-Analysis of Cotinine Levels in Cigarette Smokers Identifies Locus at 4q13.2.* Scientific reports, 2016. **6**: p. 20092.
75. Loukola, A., et al., *A Genome-Wide Association Study of a Biomarker of Nicotine Metabolism.* PLoS Genet, 2015. **11**(9): p. e1005498.
76. Hukkanen, J., P. Jacob, 3rd, and N.L. Benowitz, *Metabolism and disposition kinetics of nicotine.* Pharmacol Rev, 2005. **57**(1): p. 79-115.
77. Adams, S., *Psychopharmacology of Tobacco and Alcohol Comorbidity: a Review of Current Evidence.* Curr Addict Rep, 2017. **4**(1): p. 25-34.
78. Williams, J.M. and D. Ziedonis, *Addressing tobacco among individuals with a mental illness or an addiction.* Addict Behav, 2004. **29**(6): p. 1067-83.
79. G, S.B., et al., *Cigarette smoke and related risk factors in neurological disorders: An update.* Biomed Pharmacother, 2017. **85**: p. 79-86.
80. Hartz, S.M., et al., *Genetic correlation between smoking behaviors and schizophrenia.* Schizophr Res, 2017.
81. Manzella, F., S.E. Maloney, and G.T. Taylor, *Smoking in schizophrenic patients: A critique of the self-medication hypothesis.* World J Psychiatry, 2015. **5**(1): p. 35-46.
82. Cahill, K., et al., *Nicotine receptor partial agonists for smoking cessation.* Cochrane Database Syst Rev, 2016(5): p. CD006103.
83. Li, M.D., *The genetics of nicotine dependence.* Curr Psychiatry Rep, 2006. **8**(2): p. 158-64.
84. Loukola, A., et al., *Genetics and smoking.* Curr Addict Rep, 2014. **1**(1): p. 75-82.
85. Boardman, J.D., C.L. Blalock, and F.C. Pampel, *Trends in the genetic influences on smoking.* J Health Soc Behav, 2010. **51**(1): p. 108-23.
86. Vink, J.M., G. Willemsen, and D.I. Boomsma, *Heritability of smoking initiation and nicotine dependence.* Behav Genet, 2005. **35**(4): p. 397-406.
87. Li, M.D., et al., *A meta-analysis of estimated genetic and environmental effects on smoking behavior in male and female adult twins.* Addiction, 2003. **98**(1): p. 23-31.
88. Rose, R.J., et al., *Genetics of smoking behavior*, in *Handbook of behavior genetics*. 2009, Springer. p. 411-432.
89. Lessov-Schlaggar, C.N., et al., *Genetics of nicotine dependence and pharmacotherapy.* Biochemical pharmacology, 2008. **75**(1): p. 178-195.
90. Lessov, C.N., et al., *Defining nicotine dependence for genetic research: evidence from Australian twins.* Psychological medicine, 2004. **34**(5): p. 865-879.
91. Broms, U., et al., *Genetic architecture of smoking behavior: a study of Finnish adult twins.* Twin Research and Human Genetics, 2006. **9**(1): p. 64-72.
92. Pergadia, M.L., et al., *Nicotine withdrawal symptoms in adolescent and adult twins.* Twin Research and Human Genetics, 2010. **13**(4): p. 359-369.

93. Pergadia, M.L., et al., *Genetic analyses of DSM-IV nicotine withdrawal in adult twins.* Psychological medicine, 2006. **36**(7): p. 963-972.
94. Xian, H., et al., *The heritability of failed smoking cessation and nicotine withdrawal in twins who smoked and attempted to quit.* Nicotine & Tobacco Research, 2003. **5**(2): p. 245-254.
95. Swan, G.E., et al., *Pharmacogenetics of nicotine metabolism in twins: methods and procedures.* Twin Research and Human Genetics, 2004. **7**(5): p. 435-448.
96. Swan, G.E., et al., *Nicotine metabolism: the impact of CYP2A6 on estimates of additive genetic influence.* Pharmacogenetics and genomics, 2005. **15**(2): p. 115-125.
97. Amos, C.I., M.R. Spitz, and P. Cinciripini, *Chipping away at the genetics of smoking behavior.* Nat Genet, 2010. **42**(5): p. 366-8.
98. Tobacco and C. Genetics, *Genome-wide meta-analyses identify multiple loci associated with smoking behavior.* Nat Genet, 2010. **42**(5): p. 441-7.
99. Liu, J.Z., et al., *Meta-analysis and imputation refines the association of 15q25 with smoking quantity.* Nat Genet, 2010. **42**(5): p. 436-40.
100. Thorgeirsson, T.E., et al., *Sequence variants at CHRNB3-CHRNA6 and CYP2A6 affect smoking behavior.* Nat Genet, 2010. **42**(5): p. 448-53.
101. Siedlinski, M., et al., *Genome-wide association study of smoking behaviours in patients with COPD.* Thorax, 2011. **66**(10): p. 894-902.
102. Dempsey, D., et al., *Nicotine metabolite ratio as an index of cytochrome P450 2A6 metabolic activity.* Clin Pharmacol Ther, 2004. **76**(1): p. 64-72.
103. Rubinstein, M.L., et al., *Race, gender, and nicotine metabolism in adolescent smokers.* Nicotine Tob Res, 2013. **15**(7): p. 1311-5.
104. Lerman, C., et al., *Use of the nicotine metabolite ratio as a genetically informed biomarker of response to nicotine patch or varenicline for smoking cessation: a randomised, double-blind placebo-controlled trial.* Lancet Respir Med, 2015. **3**(2): p. 131-138.
105. Chen, L.S., et al., *Pharmacotherapy effects on smoking cessation vary with nicotine metabolism gene (CYP2A6).* Addiction, 2014. **109**(1): p. 128-137.
106. Gaedigk, A., et al., *The Pharmacogene Variation (PharmVar) Consortium: Incorporation of the Human Cytochrome P450 (CYP) Allele Nomenclature Database.* Clin Pharmacol Ther, 2018. **103**(3): p. 399-401.
107. Mwenifumbo, J.C. and R.F. Tyndale, *Molecular genetics of nicotine metabolism.* Handb Exp Pharmacol, 2009(192): p. 235-59.
108. Baurley, J.W., et al., *Genome-Wide Association of the Laboratory-Based Nicotine Metabolite Ratio in Three Ancestries.* Nicotine Tob Res, 2016. **18**(9): p. 1837-1844.
109. Chenoweth, M.J., et al., *Genome-wide association study of a nicotine metabolism biomarker in African American smokers: impact of chromosome 19 genetic influences.* Addiction, 2017.
110. Wang, J. and M.D. Li, *Common and unique biological pathways associated with smoking initiation/progression, nicotine dependence, and smoking cessation.* Neuropsychopharmacology, 2010. **35**(3): p. 702-19.
111. Minica, C.C., et al., *Pathways to smoking behaviours: biological insights from the Tobacco and Genetics Consortium meta-analysis.* Mol Psychiatry, 2017. **22**(1): p. 82-88.
112. Fisher, M.L., et al., *Role of the Neuregulin Signaling Pathway in Nicotine Dependence and Co-morbid Disorders.* Int Rev Neurobiol, 2015. **124**: p. 113-31.
113. Szklarczyk, D., et al., *The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored.* Nucleic Acids Res, 2011. **39**(Database issue): p. D561-8.
114. Loukola, A., et al., *Linkage of nicotine dependence and smoking behavior on 10q, 7q and 11p in twins with homogeneous genetic background.* Pharmacogenomics J, 2008. **8**(3): p. 209-19.

115. Loukola, A., et al., *Genome-wide association study on detailed profiles of smoking behavior and nicotine dependence in a twin sample.* Mol Psychiatry, 2014. **19**(5): p. 615-24.
116. Turner, J.R., et al., *Evidence from mouse and man for a role of neuregulin 3 in nicotine dependence.* Mol Psychiatry, 2014. **19**(7): p. 801-10.
117. Hatzimanolis, A., et al., *Multiple variants aggregate in the neuregulin signaling pathway in a subset of schizophrenia patients.* Transl Psychiatry, 2013. **3**: p. e264.
118. Savonenko, A.V., et al., *Alteration of BACE1-dependent NRG1/ErbB4 signaling and schizophrenia-like phenotypes in BACE1-null mice.* Proc Natl Acad Sci U S A, 2008. **105**(14): p. 5585-90.
119. Dejaegere, T., et al., *Deficiency of Aph1B/C-gamma-secretase disturbs Nrg1 cleavage and sensorimotor gating that can be reversed with antipsychotic treatment.* Proc Natl Acad Sci U S A, 2008. **105**(28): p. 9775-80.
120. Gao, X., et al., *DNA methylation changes of whole blood cells in response to active smoking exposure in adults: a systematic review of DNA methylation studies.* Clin Epigenetics, 2015. **7**: p. 113.
121. Lee, K.W. and Z. Pausova, *Cigarette smoking and DNA methylation.* Front Genet, 2013. **4**: p. 132.
122. Breitling, L.P., et al., *Tobacco-smoking-related differential DNA methylation: 27K discovery and replication.* Am J Hum Genet, 2011. **88**(4): p. 450-7.
123. Wan, E.S., et al., *Cigarette smoking behaviors and time since quitting are associated with differential DNA methylation across the human genome.* Hum Mol Genet, 2012. **21**(13): p. 3073-82.
124. Siedlinski, M., et al., *Association of cigarette smoking and CRP levels with DNA methylation in alpha-1 antitrypsin deficiency.* Epigenetics, 2012. **7**(7): p. 720-8.
125. Shenker, N.S., et al., *Epigenome-wide association study in the European Prospective Investigation into Cancer and Nutrition (EPIC-Turin) identifies novel genetic loci associated with smoking.* Hum Mol Genet, 2013. **22**(5): p. 843-51.
126. Philibert, R.A., S.R. Beach, and G.H. Brody, *Demethylation of the aryl hydrocarbon receptor repressor as a biomarker for nascent smokers.* Epigenetics, 2012. **7**(11): p. 1331-8.
127. Sun, Y.V., et al., *Epigenomic association analysis identifies smoking-related DNA methylation sites in African Americans.* Hum Genet, 2013. **132**(9): p. 1027-37.
128. Zeilinger, S., et al., *Tobacco smoking leads to extensive genome-wide changes in DNA methylation.* PLoS One, 2013. **8**(5): p. e63812.
129. Philibert, R.A., et al., *Changes in DNA methylation at the aryl hydrocarbon receptor repressor may be a new biomarker for smoking.* Clin Epigenetics, 2013. **5**(1): p. 19.
130. Besingi, W. and A. Johansson, *Smoke-related DNA methylation changes in the etiology of human disease.* Hum Mol Genet, 2014. **23**(9): p. 2290-7.
131. Elliott, H.R., et al., *Differences in smoking associated DNA methylation patterns in South Asians and Europeans.* Clin Epigenetics, 2014. **6**(1): p. 4.
132. Dogan, M.V., et al., *The effect of smoking on DNA methylation of peripheral blood mononuclear cells from African American women.* BMC Genomics, 2014. **15**: p. 151.
133. Tsaprouni, L.G., et al., *Cigarette smoking reduces DNA methylation levels at multiple genomic loci but the effect is partially reversible upon cessation.* Epigenetics, 2014. **9**(10): p. 1382-96.
134. Flanagan, J.M., et al., *Temporal stability and determinants of white blood cell DNA methylation in the breakthrough generations study.* Cancer Epidemiol Biomarkers Prev, 2015. **24**(1): p. 221-9.
135. Guida, F., et al., *Dynamics of smoking-induced genome-wide methylation changes with time since smoking cessation.* Hum Mol Genet, 2015. **24**(8): p. 2349-59.
136. Zaghlool, S.B., et al., *Association of DNA methylation with age, gender, and smoking in an Arab population.* Clin Epigenetics, 2015. **7**: p. 6.

137. Allione, A., et al., *Novel epigenetic changes unveiled by monozygotic twins discordant for smoking habits.* PLoS One, 2015. **10**(6): p. e0128265.
138. Qiu, W., et al., *The impact of genetic variation and cigarette smoke on DNA methylation in current and former smokers from the COPDGene study.* Epigenetics, 2015. **10**(11): p. 1064-73.
139. Sayols-Baixeras, S., et al., *Identification of a new locus and validation of previously reported loci showing differential methylation associated with smoking. The REGICOR study.* Epigenetics, 2015. **10**(12): p. 1156-65.
140. Zhang, Y., et al., *Self-reported smoking, serum cotinine, and blood DNA methylation.* Environ Res, 2016. **146**: p. 395-403.
141. Ambatipudi, S., et al., *Tobacco smoking-associated genome-wide DNA methylation changes in the EPIC study.* Epigenomics, 2016. **8**(5): p. 599-618.
142. Joehanes, R., et al., *Epigenetic Signatures of Cigarette Smoking.* Circ Cardiovasc Genet, 2016. **9**(5): p. 436-447.
143. Lee, M.K., et al., *DNA methylation and smoking in Korean adults: epigenome-wide association study.* Clin Epigenetics, 2016. **8**: p. 103.
144. Dogan, M.V., S.R.H. Beach, and R.A. Philibert, *Genetically contextual effects of smoking on genome wide DNA methylation.* Am J Med Genet B Neuropsychiatr Genet, 2017.
145. Hankinson, O., *The aryl hydrocarbon receptor complex.* Annu Rev Pharmacol Toxicol, 1995. **35**: p. 307-40.
146. Zhang, Y., et al., *F2RL3 methylation in blood DNA is a strong predictor of mortality.* Int J Epidemiol, 2014. **43**(4): p. 1215-25.
147. Teschendorff, A.E., et al., *Correlation of Smoking-Associated DNA Methylation Changes in Buccal Cells With DNA Methylation Changes in Epithelial Cancer.* JAMA Oncol, 2015. **1**(4): p. 476-85.
148. Sundar, I.K., et al., *DNA methylation profiling in peripheral lung tissues of smokers and patients with COPD.* Clin Epigenetics, 2017. **9**: p. 38.
149. Breitling, L.P., *Current genetics and epigenetics of smoking/tobacco-related cardiovascular disease.* Arterioscler Thromb Vasc Biol, 2013. **33**(7): p. 1468-72.
150. Morris, P.B., et al., *Cardiovascular Effects of Exposure to Cigarette Smoke and Electronic Cigarettes: Clinical Perspectives From the Prevention of Cardiovascular Disease Section Leadership Council and Early Career Councils of the American College of Cardiology.* J Am Coll Cardiol, 2015. **66**(12): p. 1378-91.
151. Ma, Y. and M.D. Li, *Establishment of a Strong Link Between Smoking and Cancer Pathogenesis through DNA Methylation Analysis.* Sci Rep, 2017. **7**(1): p. 1811.
152. Stueve, T.R., et al., *Epigenome-wide analysis of DNA methylation in lung tissue shows concordance with blood studies and identifies tobacco smoke-inducible enhancers.* Hum Mol Genet, 2017. **26**(15): p. 3014-3027.
153. Joubert, B.R., et al., *DNA Methylation in Newborns and Maternal Smoking in Pregnancy: Genome-wide Consortium Meta-analysis.* Am J Hum Genet, 2016. **98**(4): p. 680-96.
154. Kaprio, J., *The Finnish Twin Cohort Study: an update.* Twin Res Hum Genet, 2013. **16**(1): p. 157-62.
155. Kaprio, J., et al., *The Finnish Twin Registry: formation and compilation, questionnaire study, zygosity determination procedures, and research program.* Prog Clin Biol Res, 1978. **24 Pt B**: p. 179-84.
156. Kaprio, J. and M. Koskenvuo, *Genetic and environmental factors in complex diseases: the older Finnish Twin Cohort.* Twin Res, 2002. **5**(5): p. 358-65.
157. Kaprio, J., *Twin studies in Finland 2006.* Twin Res Hum Genet, 2006. **9**(6): p. 772-7.
158. Cannon, T.D., et al., *The genetic epidemiology of schizophrenia in a Finnish twin cohort. A population-based modeling study.* Arch Gen Psychiatry, 1998. **55**(1): p. 67-74.

159. Oresic, M., et al., *Phospholipids and insulin resistance in psychosis: a lipidomics study of twin pairs discordant for schizophrenia.* Genome Med, 2012. **4**(1): p. 1.
160. Association, A.P., *Diagnostic and statistical manual of mental disorders: DSM-IV*. 4th ed. 1994, Washington (DC): American Psychiatric Association.
161. Koskinen, S.M., et al., *A longitudinal twin study of effects of adolescent alcohol abuse on the neurophysiology of attention and orienting.* Alcohol Clin Exp Res, 2011. **35**(7): p. 1339-50.
162. Naukkarinen, J., et al., *Causes and consequences of obesity: the contribution of recent twin studies.* Int J Obes (Lond), 2012. **36**(8): p. 1017-24.
163. Kaprio, J., L. Pulkkinen, and R.J. Rose, *Genetic and environmental factors in health-related behaviors: studies on Finnish twins and twin families.* Twin Res, 2002. **5**(5): p. 366-71.
164. Raitakari, O.T., et al., *Cohort profile: the cardiovascular risk in Young Finns Study.* Int J Epidemiol, 2008. **37**(6): p. 1220-6.
165. Borodulin, K., et al., *Forty-year trends in cardiovascular risk factors in Finland.* Eur J Public Health, 2015. **25**(3): p. 539-46.
166. Inouye, M., et al., *An immune response network associated with blood lipid levels.* PLoS Genet, 2010. **6**(9): p. e1001113.
167. Inouye, M., et al., *Metabonomic, transcriptomic, and genomic variation of a population cohort.* Mol Syst Biol, 2010. **6**: p. 441.
168. Broms, U., et al., *Diurnal Evening Type is Associated with Current Smoking, Nicotine Dependence and Nicotine Intake in the Population Based National FINRISK 2007 Study.* J Addict Res Ther, 2012. **S2**.
169. Nikkanen, J., et al., *Mitochondrial DNA Replication Defects Disturb Cellular dNTP Pools and Remodel One-Carbon Metabolism.* Cell Metab, 2016. **23**(4): p. 635-48.
170. Roman-Garcia, P., et al., *Vitamin B(1)(2)-dependent taurine synthesis regulates growth and bone mass.* J Clin Invest, 2014. **124**(7): p. 2988-3002.
171. Aulchenko, Y.S., et al., *GenABEL: an R library for genome-wide association analysis.* Bioinformatics, 2007. **23**(10): p. 1294-6.
172. Benowitz, N.L., et al., *Optimal serum cotinine levels for distinguishing cigarette smokers and nonsmokers within different racial/ethnic groups in the United States between 1999 and 2004.* Am J Epidemiol, 2009. **169**(2): p. 236-48.
173. Luostarinen, M., et al., *Weight concerns among Finnish ever-smokers: a population-based study.* Nicotine Tob Res, 2013. **15**(10): p. 1696-704.
174. Broms, U., et al., *Analysis of detailed phenotype profiles reveals CHRNA5-CHRNA3-CHRNB4 gene cluster association with several nicotine dependence traits.* Nicotine Tob Res, 2012. **14**(6): p. 720-33.
175. Altman, D.G. and P. Royston, *The cost of dichotomising continuous variables.* BMJ, 2006. **332**(7549): p. 1080.
176. Bucholz, K.K., et al., *A new, semi-structured psychiatric interview for use in genetic linkage studies: a report on the reliability of the SSAGA.* J Stud Alcohol, 1994. **55**(2): p. 149-58.
177. Saccone, S.F., et al., *Genetic linkage to chromosome 22q12 for a heavy-smoking quantitative trait in two independent samples.* Am J Hum Genet, 2007. **80**(5): p. 856-66.
178. Houseman, E.A., et al., *DNA methylation arrays as surrogate measures of cell mixture distribution.* BMC Bioinformatics, 2012. **13**: p. 86.
179. Auer, P.L. and G. Lettre, *Rare variant association studies: considerations, challenges and opportunities.* Genome Med, 2015. **7**(1): p. 16.
180. Ipe, J., et al., *High-Throughput Assays to Assess the Functional Impact of Genetic Variants: A Road Towards Genomic-Driven Medicine.* Clin Transl Sci, 2017. **10**(2): p. 67-77.

181. Chen, Y.A., et al., *Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray.* Epigenetics, 2013. **8**(2): p. 203-9.
182. Price, M.E., et al., *Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array.* Epigenetics Chromatin, 2013. **6**(1): p. 4.
183. Naeem, H., et al., *Reducing the risk of false discovery enabling identification of biologically significant genome-wide methylation status using the HumanMethylation450 array.* BMC Genomics, 2014. **15**: p. 51.
184. Smyth, G.K., *Limma: linear models for microarray data*, in *Bioinformatics and computational biology solutions using R and Bioconductor*. 2005, Springer. p. 397-420.
185. Pidsley, R., et al., *A data-driven approach to preprocessing Illumina 450K methylation array data.* BMC Genomics, 2013. **14**: p. 293.
186. Zhou, W., P.W. Laird, and H. Shen, *Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes.* Nucleic Acids Res, 2017. **45**(4): p. e22.
187. Bolstad, B.M., et al., *A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.* Bioinformatics, 2003. **19**(2): p. 185-93.
188. Zheutlin, A.B., et al., *Cognitive endophenotypes inform genome-wide expression profiling in schizophrenia.* Neuropsychology, 2016. **30**(1): p. 40-52.
189. International HapMap, C., *A haplotype map of the human genome.* Nature, 2005. **437**(7063): p. 1299-320.
190. Benjamini, Y. and Y. Hochberg, *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.* Journal of the Royal Statistical Society. Series B (Methodological), 1995. **57**(1): p. 289-300.
191. Zhou, X. and M. Stephens, *Genome-wide efficient mixed-model analysis for association studies.* Nat Genet, 2012. **44**(7): p. 821-4.
192. Li, B. and S.M. Leal, *Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data.* Am J Hum Genet, 2008. **83**(3): p. 311-21.
193. Therneau, T.M. *coxme: Mixed effects Cox models. R package version 2.1-2.* 2015 15 June; Available from: http://CRAN.R-project.org/package=coxme.
194. Vazquez, A.I., et al., *Technical note: an R package for fitting generalized linear mixed models in animal breeding.* J Anim Sci, 2010. **88**(2): p. 497-504.
195. Wu, M.C., et al., *Rare-variant association testing for sequencing data with the sequence kernel association test.* Am J Hum Genet, 2011. **89**(1): p. 82-93.
196. He, L., et al., *Hierarchical Bayesian model for rare variant association analysis integrating genotype uncertainty in human sequence data.* Genet Epidemiol, 2015. **39**(2): p. 89-100.
197. Sellke, T.A. and B.J. Schneider, *The effects of reduced attrition on craniofacial and dentoalveolar development in the rat.* Angle Orthod, 1977. **47**(4): p. 313-22.
198. Stephens, M. and D.J. Balding, *Bayesian statistical methods for genetic association studies.* Nat Rev Genet, 2009. **10**(10): p. 681-90.
199. Hiekkalinna, T., et al., *PSEUDOMARKER: a powerful program for joint linkage and/or linkage disequilibrium analysis on mixtures of singletons and related individuals.* Hum Hered, 2011. **71**(4): p. 256-66.
200. Gertz, E.M., et al., *PSEUDOMARKER 2.0: efficient computation of likelihoods using NOMAD.* BMC Bioinformatics, 2014. **15**: p. 47.
201. Bates, D., et al., *Fitting Linear Mixed-Effects Models Using lme4.* 2015, 2015. **67**(1): p. 48.
202. Turner, S.D., *qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots.* bioRxiv, 2014.

203. Barrett, J.C., *Haploview: Visualization and analysis of SNP genotype data.* Cold Spring Harb Protoc, 2009. **2009**(10): p. pdb ip71.
204. Cuellar-Partida, G., M.E. Renteria, and S. MacGregor, *LocusTrack: Integrated visualization of GWAS results and genomic annotation.* Source Code Biol Med, 2015. **10**: p. 1.
205. Zhang, H., P. Meltzer, and S. Davis, *RCircos: an R package for Circos 2D track plots.* BMC Bioinformatics, 2013. **14**: p. 244.
206. Shabalin, A.A., *Matrix eQTL: ultra fast eQTL analysis via large matrix operations.* Bioinformatics, 2012. **28**(10): p. 1353-8.
207. McLaren, W., et al., *The Ensembl Variant Effect Predictor.* bioRxiv, 2016.
208. Kramer, A., et al., *Causal analysis approaches in Ingenuity Pathway Analysis.* Bioinformatics, 2014. **30**(4): p. 523-30.
209. Consortium, G.T., *Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans.* Science, 2015. **348**(6235): p. 648-60.
210. Ramasamy, A., et al., *Genetic variability in the regulation of gene expression in ten regions of the human brain.* Nat Neurosci, 2014. **17**(10): p. 1418-28.
211. Gaunt, T.R., et al., *Systematic identification of genetic influences on methylation across the human life course.* Genome Biol, 2016. **17**: p. 61.
212. Hannon, E., et al., *Methylation QTLs in the developing brain and their enrichment in schizophrenia risk loci.* Nat Neurosci, 2016. **19**(1): p. 48-54.
213. Wu, Y., Y.G. Yao, and X.J. Luo, *SZDB: A Database for Schizophrenia Genetic Research.* Schizophr Bull, 2016.
214. Ward, L.D. and M. Kellis, *HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease.* Nucleic Acids Res, 2016. **44**(D1): p. D877-81.
215. Xiong, H.Y., et al., *RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease.* Science, 2015. **347**(6218): p. 1254806.
216. Benowitz, N.L., *Nicotine addiction.* N Engl J Med, 2010. **362**(24): p. 2295-303.
217. Chenoweth, M.J., et al., *CYP2A6 slow nicotine metabolism is associated with increased quitting by adolescent smokers.* Pharmacogenet Genomics, 2013. **23**(4): p. 232-5.
218. Saccone, N.L., et al., *The Value of Biosamples in Smoking Cessation Trials: A Review of Genetic, Metabolomic, and Epigenetic Findings.* Nicotine Tob Res, 2017.
219. Chenoweth, M.J., et al., *Known and novel sources of variability in the nicotine metabolite ratio in a large sample of treatment-seeking smokers.* Cancer Epidemiol Biomarkers Prev, 2014. **23**(9): p. 1773-82.
220. Connor Gorber, S., et al., *The accuracy of self-reported smoking: a systematic review of the relationship between self-reported and cotinine-assessed smoking status.* Nicotine Tob Res, 2009. **11**(1): p. 12-24.
221. Su, D., et al., *Distinct Epigenetic Effects of Tobacco Smoking in Whole Blood and among Leukocyte Subtypes.* PLoS One, 2016. **11**(12): p. e0166486.
222. Zanger, U.M. and M. Schwab, *Cytochrome P450 enzymes in drug metabolism: regulation of gene expression, enzyme activities, and impact of genetic variation.* Pharmacol Ther, 2013. **138**(1): p. 103-41.
223. Drgon, T., et al., *Genome-wide association for nicotine dependence and smoking cessation success in NIH research volunteers.* Mol Med, 2009. **15**(1-2): p. 21-7.
224. Sung, Y.J., et al., *Gene-smoking interactions identify several novel blood pressure loci in the Framingham Heart Study.* Am J Hypertens, 2015. **28**(3): p. 343-54.
225. Uhl, G.R., et al., *Molecular genetics of successful smoking cessation: convergent genome-wide association study results.* Arch Gen Psychiatry, 2008. **65**(6): p. 683-93.

226. Repapi, E., et al., *Genome-wide association study identifies five loci associated with lung function.* Nat Genet, 2010. **42**(1): p. 36-44.
227. Pesce, A., et al., *An evaluation of the diagnostic accuracy of liquid chromatography-tandem mass spectrometry versus immunoassay drug testing in pain patients.* Pain Physician, 2010. **13**(3): p. 273-81.
228. Agrawal, A., et al., *An autosomal linkage scan for cannabis use disorders in the nicotine addiction genetics project.* Arch Gen Psychiatry, 2008. **65**(6): p. 713-21.
229. Karlsson Linner, R., et al., *An epigenome-wide association study meta-analysis of educational attainment.* Mol Psychiatry, 2017.
230. Johnson, W., et al., *Does education confer a culture of healthy behavior? Smoking and drinking patterns in Danish twins.* Am J Epidemiol, 2011. **173**(1): p. 55-63.
231. Okbay, A., et al., *Genome-wide association study identifies 74 loci associated with educational attainment.* Nature, 2016. **533**(7604): p. 539-42.
232. Hiekkalinna, T., et al., *On the statistical properties of family-based association tests in datasets containing both pedigrees and unrelated case-control samples.* Eur J Hum Genet, 2012. **20**(2): p. 217-23.
233. Rieger, J.K., et al., *Expression variability of absorption, distribution, metabolism, excretion-related microRNAs in human liver: influence of nongenetic factors and association with gene expression.* Drug Metab Dispos, 2013. **41**(10): p. 1752-62.
234. Zhang, Y., et al., *Comparison and combination of blood DNA methylation at smoking-associated genes and at lung cancer-related genes in prediction of lung cancer mortality.* Int J Cancer, 2016. **139**(11): p. 2482-92.
235. Marabita, F., et al., *Smoking induces DNA methylation changes in Multiple Sclerosis patients with exposure-response relationship.* Sci Rep, 2017. **7**(1): p. 14589.
236. Zhang, Y., et al., *DNA methylation signatures in peripheral blood strongly predict all-cause mortality.* Nat Commun, 2017. **8**: p. 14617.
237. Ainsworth, H.F., S.Y. Shin, and H.J. Cordell, *A comparison of methods for inferring causal relationships between genotype and phenotype using additional biological measurements.* Genet Epidemiol, 2017.
238. Teschendorff, A.E. and C.L. Relton, *Statistical and integrative system-level analysis of DNA methylation data.* Nat Rev Genet, 2017.
239. Nelson, S.C., et al., *Imputation-based genomic coverage assessments of current human genotyping arrays.* G3 (Bethesda), 2013. **3**(10): p. 1795-807.
240. Surakka, I., et al., *The rate of false polymorphisms introduced when imputing genotypes from global imputation panels.* bioRxiv, 2016.
241. Peltonen, L., A. Jalanko, and T. Varilo, *Molecular genetics of the Finnish disease heritage.* Hum Mol Genet, 1999. **8**(10): p. 1913-23.
242. Bloom, A.J., et al., *Use of a predictive model derived from in vivo endophenotype measurements to demonstrate associations with a complex locus, CYP2A6.* Hum Mol Genet, 2012. **21**(13): p. 3050-62.
243. Bloom, J., et al., *The contribution of common CYP2A6 alleles to variation in nicotine metabolism among European-Americans.* Pharmacogenet Genomics, 2011. **21**(7): p. 403-16.
244. King, D.P., et al., *Smoking cessation pharmacogenetics: analysis of varenicline and bupropion in placebo-controlled clinical trials.* Neuropsychopharmacology, 2012. **37**(3): p. 641-50.
245. Papaleo, F., et al., *Behavioral, Neurophysiological, and Synaptic Impairment in a Transgenic Neuregulin1 (NRG1-IV) Murine Schizophrenia Model.* J Neurosci, 2016. **36**(17): p. 4859-75.
246. Rico, B., *Finding a druggable target for schizophrenia.* Proc Natl Acad Sci U S A, 2012. **109**(30): p. 11902-3.

247. Tang, J., et al., *DNA methylation and personalized medicine.* J Clin Pharm Ther, 2014. **39**(6): p. 621-7.
248. Purcell, S., S.S. Cherny, and P.C. Sham, *Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits.* Bioinformatics, 2003. **19**(1): p. 149-50.
249. Faul, F., et al., *Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses.* Behav Res Methods, 2009. **41**(4): p. 1149-60.